AFHRL-TR-79-78

LEVEL

AD A077347

19th Annual Conference of the Military Testing Association

PROCEEDINGS

NOV 20 1979

MTA

THE MILITARY TESTING ASSOCIATION

Coordinated Jointly By

Air Force Human Resources Laboratory

and

U S A F Occupational Measurement Center

San Antonio, Texas

17–21 October 1977

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFHRL-TR-79-78	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) 19TH ANNUAL CONFERENCE OF THE MILITARY TESTING ASSOCIATION		5. TYPE OF REPORT & PERIOD COVERED Final
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62703F 7734 7719
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE November 1979
		13. NUMBER OF PAGES 1458
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Proceedings of the 19th Annual Conference of the Military Testing Association, San Antonio, Texas, 17–21 October 1977

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Armed Services Vocational Aptitude Battery (ASVAB)
instructional systems development
job evaluation
knowledge tests
management information system
organizational assessment
performance appraisal
personnel testing
pilot selection
team effectiveness

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Military Testing Association Conference was held 17–21 October 1977 at the El Tropicano Hotel, San Antonio, Texas. It was hosted by the Air Force Human Resources Laboratory. Independent presentations were made by members of the Department of Defense, United States Coast Guard, and related Defense contractors. The text of each presentation has been included in this report.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

NOTICE

This final report was submitted by Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235, under projects 7734 and 7719.

This technical report has been reviewed and is approved for publication.

RONALD W. TERRY, Colonel, USAF
Commander

19TH ANNUAL CONFERENCE

OF THE

MILITARY TESTING ASSOCIATION

MTA PRESIDENTS

COLONEL DAN D. FULGHAM, USAF, AFHRL

COLONEL JAMES A. TURNER, USAF, OMC

CONFERENCE COMMITTEE

DR WALTER E. DRISKILL

MS JOYCE GIORGIA

MR R. BRUCE GOULD

DR JOE T. HAZEL

COL RALPH S. HOGGATT (CHAIRMAN)

MS CECILIA KNUTSON

MS NANCY LEWIS

MAJ JIMMY L. MITCHELL

COORDINATED JOINTLY BY

AIR FORCE HUMAN RESOURCES LABORATORY

AND

USAF OCCUPATIONAL MEASUREMENT CENTER

SAN ANTONIO, TEXAS

17-21 OCTOBER 1977

FOREWORD

THE PAPERS PRESENTED AT THE NINETEENTH ANNUAL CONFERENCE OF THE MILITARY TESTING ASSCCIATION CAME FROM THE BUSINESS, EDUCATIONAL, AND MILITARY COMMUNITIES, BOTH FOREIGN AND DOMESTIC. THE PAPERS REFLECT THE OPINIONS OF THEIR AUTHORS ONLY AND ARE NOT TO BE CONSTRUED AS THE OFFICIAL POLICY OF ANY INSTITUTION, GOVERNMENT, OR ARMED-SERVICE BRANCH.

TABLE OF CONTENTS

Page

Page

Page

Page

Introductory Remarks

Col James A. Turner
USAF Occupational Measurement Center

Good morning. Welcome to San Antonio, one of America's four unique cities, home of the Air Force Human Resources Laboratory, the Occupational Measurement Center, and Alamo Chapter USAF. It is our pleasure to extend a welcome to our keynote speaker, General Emanuel, and to our distinguished service representatives of the Canadian, Australian, and German Forces and the U.S. Army, Navy, Coast Guard, and Air Force. On behalf of all the staff of the Human Resources Lab, who led the way; the Occupational Measurement Center, who helped; and particularly from Col Dan G. Fulgham, Commander of AFHRL (who had to be in Washington today), I give you all greetings and welcome. We hope you enjoy your stay here in San Antonio.

We open this conference at a time when there is considerable hue and cry and controversy over testing, particularly intelligence testing, and the use of scores in educational placement. It is appropriate for us to take a frequent look at what we do with testing and the fairness of the instruments we use. Our interest cannot be limited to testing alone since building good tests goes well beyond just the construction of fair and equitable examinations. We must know the tasks, the skills, the knowledges required if the individial is to perform satisfactorily. Once we know these things, then one can proceed with how to train the individual, and we can determine the areas which are most realistically tested to measure the person's promotion or other potential. In short, testing is a major component in development and maintenance of the Air Force Personnel System.

Our guest of honor has had a good deal to do with the AF system and is well qualified to keynote this 19th MTA conference. General Herbert L. Emanuel, USAF, is Deputy to the Assistant Deputy Chief of Staff, Personnel for Military Personnel, and Vice Commander of the Air Force Military Personnel Center at Randolph Air Force Base, Texas. He has a Baccalaureate in Communicative Arts, a Master's Degree in Personnel Management, and is a graduate of the Armed Forces Staff College and the Industrial College of the Armed Forces. General Emanuel entered the Air Force as a 2nd Lt via the ROTC program in May 1952 and served three successive assignments as a personnel officer. After many enriching and broadening assignments worldwide, including a stint as an Information Officer, Director of Cadet Activities at the Air Force Academy, and a tour in Vietnam, he was assigned to HQ USAF as a Personnel Planner. He was involved in the initial publication of the USAF Personnel Plan and development of the AF Personnel Management by Objectives System, which still guides Air Force personnel activities. After completing the Industrial College of the Armed Forces, he returned to the Headquarters. As Deputy Director of Personnel for Plans and Policy, he was instrumental in the design and implementation of many initiatives calculated to foster morale and force discipline in the austere personnel environment of the 1970's. He assumed his present position in March 1976. It is with great pleasure I introduce to you General Herb Emanuel.

FUTURE TRENDS IN MILITARY PERSONNEL
TESTING: SOME SPECULATIONS

BY

BRIG GEN H. L. EMANUEL
AIR FORCE MILITARY PERSONNEL CENTER

Distinguished Guests, Testing Researchers, Testing Practitioners, Personnel Managers, Members of the Military Testing Association.

It is certainly a pleasure and a privilege to speak to you today. I was particularly pleased to learn that your membership includes representatives from the Canadian Armed Forces and The Royal Australian Air Force, in addition to representatives from the US military services, academia, and the business world. The important work which you do should certainly be enhanced by the spirit of international cooperation which you enjoy.

Within the military, we are confronted with the same personnel problems as any other organization, whether large or small, public or private--that of shaping and adapting available human resources into useful and effective manpower. The very multiplicity of skills required by the military poses problems in personnel, training, and manpower utilization which are unprecedented. Personnel requirements change rapidly and on a large scale, and are dependent to a large extent upon technological advances and the international political situation.

Obviously, military personnel management is a highly complex affair. As you know, to cope with these complexities requires creative and innovative personnel research--research which addresses all aspects of the personnel life cycle: Selection, classification, training, performance appraisal, promotion, and organizational development. All of these involve test instruments of some kind. Such topics are of great interest to us in the personnel management business--an interest engendered from two basic sources. First, we are users of your product. Our effectiveness as personnel managers hinges on the successful application of techniques and procedures developed from past personnel research. Second, we are sponsors of your research. In that role, we serve as the liaison agency between you and military functional managers outside the personnel community--encouraging, explaining, and extolling the virtues of research and its applications.

Thus, we have a very close and empethetic relationship with personnel research scientists. We depend on you for timely and efficient solutions to management problems as well as for input into the formulation of personnel policy. You, in turn, depend on us as sort of public relations experts who insure your various efforts are understood and appreciated not only across the military rank and file, but at the highest echelons of service and defense management as well.

Now, to the subject at hand. At the 1973 MTA Conference, General John W. Roberts, then Air Force Deputy Chief of Staff for Personnel, proposed several refinements needed if future military personnel tests were to make significant contributions in the improvement of military personnel management. Those suggestions were subsequently adopted as objectives/standards for the development of future Air Force testing. Let me summarize them for you:

First, tests of the future should focus more on the individual--on his unique talents and desires.

Second, tests should maximize the opportunity for choice--both by the individual and by his employer.

Third--as a function of the first two--tests should provide broader profiles of information.

Since the 1973 Conference, we have continued our efforts to streamline our personnel management techniques to make more effective use of our available human talent. These efforts have included new and more sophisticated ways to assess the aptitudes and attitudes of our people. Since this is the first time the Air Force has hosted the MTA since 1973, it now seems particularly appropriate to report on some of our ongoing testing projects and see how well they meet the standards established at that time. In so doing, I believe that trends for future testing will begin to emerge.

Computerized Adaptive Testing

Computerized Adaptive Testing represents the first real potential breakthrough in the personnel testing area in the last 25 years. Although the most noticeable change in the new method of testing is the fact that the test is administered by computer, the essential difference between this method and paper-and-pencil tests is that each examinee will answer a special set of test questions "tailored" to his/her ability. Adaptive testing is a way of allowing

those tested to answer only those questions that are suited to their individual ability. This contrasts with conventional group testing procedures which require many people to spend time on questions that are either too easy or difficult for them.

Computerized adaptive testing can have major benefits, both in efficiency and test quality. A test can be taken at any time; no examiner time or special scheduling is needed. The examination time will be shorter; several abilities can be tested in the time it now takes for one. And, because examinees cannot be sure which questions will be asked, it retards, if not eliminates, the problem of test compromise.

Air Force research in this area is directed toward possible application at the 66 nation-wide armed forces examining and entrance stations (AFEES). The Air Force Human Resources Laboratory (AFHRL) has already prepared a prototype demonstration model which is currently on-line at the San Antonio AFEES. (In fact, this prototype is available here at the conference should you care to see it.) In addition to providing personnel managers with a look at what computerized testing is all about, it is enabling AFHRL to gain first hand knowledge of AFEES requirements vis-a-vis computer arrays so that future hardware may be more appropriately human engineered.

Obviously, before computers can be used to test applicants for military service, the cost of procuring computer systems, display terminals, and related technology will have to be substantially reduced. Yet, we know that this is the enlistment testing of the future, so we have started planning for it now. In that regard, a joint service working group on R&D applications of computer technology to military personnel acquisitions has been formed to oversee and coordinate work in each of the services. In addition to R&D, implementation poses unique logistical/managerial problems which need to be addressed. The U.S. Civil Service Commission has set up a task force to consider those kinds of issues. We trust they will share their experiences with us. In any event, it behooves us to look at all aspects of testing by computer so when the "Science" is finally ready, so is "Management."

Prediction of Motivational Attritions

The high rate of involuntary attrition that occurs among military personnel is the subject of growing concern

at all levels of the Department of Defense (DOD). The Defense Manpower Commission in a recent report has noted that DOD incurs an annual cost of approximately one billion dollars because one out of every four DOD accessions is involuntarily separated prior to completion of the first term of enlistment. A great percentage of those discharged are identified by the training centers in the early stages of the enlistee's basic or recruit training. Each of the services operates a program designed to identify as early as possible those who will ultimately fail to adapt to the military service and to separate them administratively.

The use of the early discharge programs as a screening device is both inefficient and costly, compared to screening programs operated at the point of entry, prior to enlistment. Thus, one solution to the problem is to increase the effectiveness of the pre-enlistment selection system so as to better predict the probability of an individual's successful adaptation to the military life.

One effort to solve this problem which is currently ongoing is the Motivational Attrition Prediction (MAP) model. This new approach applies the maximum likelihood estimation technique which we believe will achieve better differentiation between potential failures and successes.

The model was initially tested at the United States Air Force Academy. To evaluate the method in an operational setting, i.e., as a screening device, a prediction equation was developed using the Class of 1977 and applied a priori to the Class of 1979. Within six months, 49% of the predicted failures had resigned.

Because results of the Air Force Academy study were so promising, the model was next tested using 15,000 1972 Air Force enlistees to predict their first term involuntary attrition. After estimating and applying a new predition equation, 57% of the group predicted to fail had been involuntarily discharged prior to completing their enlistment.

Such success with the MAP model has prompted still further research into its utility as a pre-enlistment screening tool. Beginning in May 1977, a comprehensive joint service validation study was initiated. Data, including aptitudinal and biographical information, were collected on more than 70,000 applicants for all military services. Predicted attritions based on the equation

derived from the 1972 Air Force sample will first be compared against actual service attritions to evaluate the accuracy of the model. Refinement of that equation, as well as development of ones specific for each service, will also be accomplished as the attrition data further matures.

If the MAP model works as well operationally as it did experimentally with the 1972 Air Force recruits, all services can benefit. The higher retention caused by selecting recruits more likely to complete their first enlistment would save not only in training costs but would also enhance recruiting.

Literacy and the Measurement of Reading

Recently, the General Accounting Office (GAO) has submitted a report on illiteracy in the military services to the Secretary of Defense and recommended, among other things, that DOD have the services establish an overall minimum reading level required for enlistment and determine the reading grade level required for each military occupation. In addition, the Congress has become concerned about the problem of the services' attempting to correct educational deficiencies of enlistees after they enter active duty. Congress has suggested that perhaps a more efficient approach would be for potential enlistees with educational weaknesses to receive basic skills training prior to enlistment. Accordingly, the Secretaries of Health, Education and Welfare; and Labor, in coordination with the Secretary of Defense, have been requested to develop such a basic skills program. Of course, one aspect of this type of education would be remedial reading.

Clearly, the next step in attacking the literacy problem is to include reading comprehension as one of our criteria for enlistment eligibility. Currently, we have no direct measure of reading ability--only an approximation derived from our aptitude battery. As a result, individuals with literacy problems are not identified until after they experience academic or job performance difficulties. Obviously, both the Congress and GAO believe something should be done to alleviate this service-wide situation.

As a first step, we are developing our own reading test; if it proves valid and equitable for all groups, we plan to consider its use as a screening device to

select out individuals who have inadequate reading skills. Moreover, since GAO has recommended such an approach for DOD implementation, there is the very real possibility that a reading test may be incorporated as part of ASVAB at some future date. Then, as we gain experience in determining the literacy requirements of military occupations, we could also attempt to match the reading skills of personnel (as measured by the reading test) to the reading demands of our jobs.

This discussion of testing initiatives within the Air Force is by no means a full and comprehensive one. Other efforts which merit attention, such as development and validation of vocational interest measures, pre-enlistment aptitude tests, non-verbal aptitude tests, and perceptual/psychomotor devices, were omitted. However, my purpose in selecting those mentioned was to present them as illustrations of current attempts to improve our testing programs.

When I began this morning, I promised to describe some of our ongoing testing projects and to see how well they met the Air Force objectives for future military tests. So far, I have fulfilled the first part of that promise; now, let me turn to the second. By way of brief review, the objectives mentioned earlier indicate that future tests should (a) focus on the talents and desires of the individual, (b) help maximize the opportunity for both employer and employee choice, and (c) as a function of the first two, provide broader profiles of individuals.

Well, computerized adaptive testing and literacy assessment certainly focus on the talents of individuals. A computerized test will bring additional precision to the testing situation while the reading test will assess a previous untapped skill. In addition, the vocational interest test and the perceptual/psychomotor devices also support this objective.

Under the second criterion, maximizing employee/employer choice, it seems that all the previously discussed testing techniques qualify. Improved motivational attrition prediction will allow us to screen out more of those individuals who can't make a successful adjustment to service life. In addition, reading assessment will permit individuals to enter occupations for which they possess adequate literacy skills. Then too, vocational interest measurement will give examinees the

opportunity to discover those jobs with which they will most likely be satisfied. So, it seems that both the individual and the Air Force will be happier with the addition of these types of tests to the personnel test inventory.

Finally, by virtue of satisfying the first two objectives, the third one also seems fulfilled. Certainley, more and more information about individuals will be added into the personnel selection and placement process. Thus, when evaluated against our standards for future tests, it would appear that we are moving in the right direction.

In conclusion, I hope I have conveyed my enthusiasm for the future of military personnel testing. Today, we in personnel management are facing problems we have never seen before. We have new kinds of people with differing education levels, skills, values, ambitions, and life styles that we must consider and make part of the military family. To do this requires constantly pushing the testing state-of-the-art. We must move away from the aptitude measurement-only type of testing and toward a broader assessment of other relevant dimensions of human behavior. Obviously, to perfect such new testing techniques won't be easy. It's good to be present at this conference and to know there are the kinds of people represented here who are dedicated to solving this problem.

One final comment now, if I may. As you conduct your deliberations this week, I hope you will reflect on and take pride in what the past has accomplished. However, the challenge I would impart to you this morning is to think ahead and take all possible actions to maximize the contributions of military testing to the personnel management of the future.

Thank you very much.

THE DEVELOPMENT OF A PERFORMANCE APPRAISAL SYSTEM FOR THE U.S. COAST GUARD[1]

Edwin T. Cornelius III
Milton D. Hakel
The Ohio State University

Joseph J. Cowan
Headquarters, United States Coast Guard

What I would like to do in the next forty-five minutes is share with you some of the unique aspects of a year-long study to develop an improved performance evaluation system for enlisted personnel in the United States Coast Guard. Since the project involved developing an improved performance evaluation system, I would like to take some time at the beginning of the talk to briefly describe the existing enlisted performance evaluation system in the Coast Guard.

All enlisted personnel are rated twice a year. Regardless of type of job rating or level of responsibility (rank), all personnel are evaluated on the same rating form. Three attributes of individuals are evaluated: Proficiency, Leadership, and Conduct. As Figure 1 illustrates, the existing system suffers from the usual performance rating problems inherent in a large bureaucracy. First, there is an overall inflation of marks in the system. The average rating score using the enlisted form is supposed to be 3.3, i.e., if an individual is performing in a capable and dependable fashion, and is the type of person the Coast Guard will promote on schedule, he is supposed to be rated 3.3. As Figure 1 illustrates, the average evaluation is far higher than 3.3.

A second conclusion obtained from Figure 1 is the dramatic grade effects. That is, E7's on the average are rated higher than E6's, and E6's in turn are rated higher than E5's, etc. This is true in spite of specific directions in the existing system for raters to evaluate individuals in comparison with others with the same rank and length of service. In theory at least, the average performing E9 should be rated with the same value (3.3) as the average performing E3.

A third observation is the redundancy of information in the current system. As you can see from Figure 1 the pattern of scores for Leadership and Proficiency are identical. In fact, the Pearson correlation for these data are $r = .90$. Of course this means that if you know an individual's score on the Leadership variable, you can almost perfectly predict his score for Proficiency and vice-versa. Since there is no

[1]Paper presented at the 19th Annual Military Testing Association Conference, San Antonio, October 18, 1977.

variance in the Conduct scores, essentially one piece of information about an individual is captured and communicated using the present system.

Table 1 illustrates the effect that the operating characteristics of the present system have had on administrative uses for data of this sort. One use of performance appraisal data in the Coast Guard is as an aid in making promotion decisions. Components of the promotion system include scores on a paper and pencil exam, supervisory ratings obtained through the current performance evaluation system, length of time in the service, time in grade, and medals and awards. As you can see by looking at this table, the effect of inflation of marks coupled with the large between-grade variance relative to the within grade variance has had the effect of deteriorating the contribution of performance evaluation data by a factor of 50 percent. Promotions in the Coast Guard today are chiefly determined by scores on the paper and pencil tests and length of time in service. The contribution of actual performance on the job as evaluated by supervisors is minimal.

This is the backdrop against which our project started one year ago. A lot has transpired in that year, and I obviously can't give you all the details of the project in the time allotted to me this morning. Instead, I would like to share with you some of the aspects of this project that we find different and exciting. These are characteristics of the project that we think represent something unusual either philosophically or methodologically for studies of this sort.

1. Philosophy of the Project

We have had the philosophy from the very beginning that the development of a technically perfect rating instrument by itself would not lead to good performance appraisal data for the Coast Guard. Despite the fact that you develop a technically sound instrument based on careful job analysis, and despite the fact that it has high user acceptability and is administratively simple to operate, you can still fail at this business. You can fail because what counts most in obtaining good performance appraisal information has very little to do directly with the format and psychometric characteristics of the rating instrument. What counts most is the motivation of the raters in the system to rate accurately. In this regard, an important task during the last year has been to develop a rater feedback system that can be used to build trust in the operating characteristics of the marking system.

During the course of this project we spent many hours discussing performance appraisal problems with raters while conducting technical conferences at field locations. A common sentiment expressed by

participants in all these meetings was the sincere desire to rate subordinates fairly and accurately. The officers and Chief Petty Officers that we talked to seemed to understand the need for accurate performance data for manpower planning and development purposes. The reason these raters rate leniently has nothing to do with the dynamics of having to face their men on a day in and day out basis, or any of the several other explanations for lenient ratings proposed by Bass, Glickman, Kipnis and others through the years. The reason these raters rate leniently is that they don't trust the system. "I'm willing to rate accurately, but I don't trust the other raters in the system to do the same." In this regard I discovered at one conference an elaborate informal system used to compare marks in an effort to provide informal guidelines for determining the degree of inflation to be used in evaluating subordinates.

There is no reason that an approved formal system can't provide raters with data regarding the distributions of evaluation marks in the system. All the raters we talked to agreed that if they were told how other raters in the Coast Guard were rating, and if the others were using the system properly, they would no longer rate leniently. There was almost unanimous support for a rater feedback system.

Regardless of the type of forms that were developed then, a requirement for this project was to develop a feedback system to major commands and to individual raters in the field in order to maintain trust and openness in the performance appraisal system.

2. Job Analysis Approach

The second aspect of this project that I would like to talk about is the approach that we used to study the enlisted jobs in the Coast Guard. When we started this project one of our most difficult problems was to determine the number of appraisal forms that should be developed. We were posed with the problem of having almost 30 different job ratings and 9 different ranks to study. A major activity was to determine how the different jobs and ranks could be collapsed into major groupings for which separate appraisal instruments could be developed. Once this decision was made, our task was to develop prototype forms that were user-acceptable, technically excellent, and easy to administer. In addition, the whole system had to be printed on the front and back of a single page. And, oh yes, all this was to be accomplished within a 10-month framework.

You can see that an immediate problem was how to quickly collect job analysis information from a variety of different jobs. The usual philosophy for studying jobs in the military has been the task-oriented philosophy advocated through the years by Morsh, Christal, Driskall

and others in the Air Force. However, task statements were only available on roughly half the enlisted jobs in the Coast Guard. We, therefore, decided to adopt a different type of job analysis philosophy: the worker-oriented philosophy advocated by McCormick and his associates over the years at Purdue University.

As you know, the worker-oriented approach seeks to describe jobs in terms of a limited number of universal job elements that focus on the generalized human behaviors required for work rather than work activities specific to an individual job. We thought this particularly philosophy was well suited to answer the question of how to quickly compare the large numbers of jobs and ranks in the Coast Guard. Our decision, then, was to develop a single worker oriented questionnaire specifically for the Coast Guard that could be mailed to representatives of all the different job ratings and ranks of enlisted personnel.

As a starting point in this endeavor we borrowed heavily from the most famous (and only) worker oriented questionnaire: the Position Analysis Questionnaire, developed by McCormick under another government contract. We made several changes in the PAQ to adapt it for Coast Guard use. First, we deleted items that were not at all appropriate to the military setting. Of those items that we did keep, we changed the wording and the examples to fit more clearly the Coast Guard setting. A major effort involved reducing the reading level of the PAQ. Previous research with the PAQ had shown that it required a post college graduate reading level. This is fine for trained job analysts, but would not do in a mass mail-out to enlisted personnel in the Coast Guard where the average education level was at the 12th grade level or below. In this regard we were successful in reducing the reading level from the 17th grade level to the 10th grade level, as measured by a computer program for that purpose that we have at Ohio State.

Another significant change involved eliminating the variety of different response scale formats used on the PAQ We converted all items so that they could be evaluated by the Relative Time Spent scale used in the Air Force and other military services. A final revision was to add some fifty items that we called "leadership process items." In our study of the PAQ we found that as a job analysis instrument it was particularly deficient in the area of differentiating among higher level leadership processes. Our source for these additional supervisory-type items was verb lists from previous task analyses performed in the Coast Guard.

We finally ended up with a 12-page, 153-element booklet that was mailed to some 3000 Coast Guard enlisted personnel. The sample was equally represented by members in all job ratings and grades. Incidentally,

we got a 64 percent return rate at our cutoff point. Our final sample consisted of responses from 2,023 individuals.

3. Statistical Methodology for Analyzing the Job Analysis Data

Let me now describe the data base that the returned questionnaires provided us. First of all, the data formed a three-dimensional cube. One facet of the cube had 153 levels representing the 153 different worker-oriented job elements on our questionnaire. A second facet of the cube contained 23 levels representing the 23 job ratings in the Coast Guard for which sufficient numbers of persons existed to be included in the statistical analyses. The last facet of the cube contained 9 levels representing the nine different ranks in the Coast Guard. Each cell of this three-dimensional matrix represented a unique combination of job rating, rank, and job element, and contained from 15-35 observations, depending upon the return rate for that cell. Mean relative time spent values were computed across all observations in each cell to produce a final 153 x 23 x 9 data matrix containing mean values.

The major data analysis question facing us was how to analyze simultaneously all facets of this cube and come up with practical suggestions for the number of forms that should be developed for Coast Guard use. In this regard, Tucker's three-mode factor analysis is uniquely designed to analyze data of this sort. Three-mode factor analysis proceeds in two stages. During the first stage a separate factor analysis is computed on the separate modes of the data (in our case: job elements, job ratings, and ranks). In the second stage a core matrix is created that interrelates factors from the various modes of the data.

To illustrate the kinds of meaningful output that this procedure gave us, I have included the results of the job grade factor analysis in Table 2. As you can see there were two factors that were extracted and rotated. E4's and E5's had principal loadings on Factor I, E7's E8's, and E9's had principal loadings on Factor II, and E6's had loadings on both Factor I and Factor II. In general, this told us that in terms of relative time spent on these worker oriented items, Chief Petty Officers roughly had the same pattern of responses. Likewise, E4's and E5's could be characterized as similar. E6's, however, were found to be similar to both groups. That is, some of the processes E6's have to exhibit on the job are similar to Petty Officers and some are similar to Chief Petty Officers.

The analysis of the remaining two modes produced equally interpretable results. For example, the factor analysis of the job rating mode indicated that there were five factors: one representing the various electronics

ratings, one representing the aviation ratings, one representing the deck/watch ratings (e.g., boatswain's mate), one representing the engineering ratings, and a final factor made up of service-type ratings such as musician, photojournalist, and hospital corpsman.

An inspection of the core matrix provided the final suggestion as to which groups of job ratings and ranks could be combined. From the entries in the core matrix we concluded that regardless of which job rating they came from, the responses for Chief Petty Officers were similar enough that they formed a group by themselves. Likewise, we found that there were five identifiable groups of Petty Officers, one for each of the five factors that I described to you.

These results lead us to recommend to the Coast Guard that they implement a performance appraisal system that contains seven forms. One form for Chief Petty Officers, five forms for the different types of Petty Officers, and one form for the non-rated personnel (Seamen, Airmen, and Firemen in the E1 - E3 ranks). We feel that a system of seven forms would be maximally sensitive to the different types of work and levels of responsibility inherent in the enlisted personnel population of the Coast Guard.

We took these suggestions with us to the technical conferences in the field and essentially received support from raters for a system of this sort. The whole process still amazes me! We started out with worker oriented job elements, analyzed them with a very complex multivariate statistical technique, and ended up with suggestions that were practical and acceptable to people in the field.

4. Emphasis on User Acceptance

Another important aspect of this project has been our heavy emphasis on developing a system with high user credibility and acceptability. Last spring we held several formal technical conferences with representatives from six different groups for which these forms were being developed. These meetings lasted one day each and were held at Governor's Island in New York harbor and at Elizabeth City, North Carolina. These sessions were characterized by group exercises, structured questionnaires, and open-ended discussions.

As an illustration of how we used suggestions from the field to shape the final format of these rating instruments, I would like to show you the response scale that we have included in the final proposed versions of the evaluation instruments (see Table 3). As you can see, we decided on a response scale with five categories. As an aside, you are probably aware of the controversy in the literature regarding the number of discriminations that humans can

reliably make when rating the performance of others. In terms of the optimum number of scale points, the most often cited recommendation is seven, probably based on Miller's 1956 paper on information processing capacities of humans. However, some researchers have argued that the more points the more reliable the data, and others have argued that the more scale points the less reliable the data. In reviewing the literature in this area, we were persuaded by a monte carlo study by Lissitz and Green at the University of Georgia. These researchers demonstrated that, in general, the more points on the response scale the more reliable the data. However, the level of the curve beyond five points was such that no practical increase in reliability can be achieved beyond five points on a scale.

Therefore, when we went to the field sessions last spring our bias was five points. However, our mock-up for these sessions contained a response format with eight response categories. To our surprise, we found an overwhelming preference for a rating scale with four or five points rather than one with a larger number of points as in our mock-up. All raters in the field said that they could confidently identify the extreme outlayers (outstanding performers and unsatisfactory performers). In addition, most raters said that of those that were left they could probably reliably make distinctions among three groups of personnel, roughly corresponding to average, above average, and below average performance.

We tried to get some consensus about how to label these five categories, but found it difficult. Everyone agreed to the labels "outstanding" and "unsatisfactory" for the two extreme rating categories on eigher end of the scale. However, there was no strong indication of how to label the three categories in the middle of the scale in a way that would convey the same meaning to all raters in the field. It was finally suggested that regardless of what the middle three categories were labeled, the most helpful information would be an indication of the suggested distribution of ratees that should fall in these middle categories. As you can see in Table 3 all three boxes are simply labeled "good," and the values 10%-70%-10% have been included on the final format. Raters in the field felt that the values 10-70-10 would interpret and give meaning to what was meant by the three levels of "good" performance. These values (10-70-10), incidentally, reflect to a great extent what the raters in the field felt the actual distribution of talent in the Coast Guard was. That is, most raters believed that the overwhelming majority of enlisted personnel in the Coast Guard were doing a good, capable job. In addition, these raters felt that a much smaller percentage of Coast Guard personnel performed a little better and a little worse than this majority. And, finally, an extremely small percentage (5%) were either outstanding or unsatisfactory performers.

Decisions about other characteristics of the final evaluation forms were also made during these field conferences. The final seven forms in the system, for example, each have two major blocks of rating items: those measuring personal qualities (e.g., Dependability, Initiative) and those measuring performance of duties. The list of personal qualities to be rated are constant across all forms. That is, we felt that the personal attributes important for success as a Petty Officer Boatswain's Mate are the same as the personal attributes necessary for a Petty Officer Radioman. However, the performance of duties items were for the most part unique to each form. These items were selected from the three-mode factor analysis output and tended to be items that had high relative time spent ratings and at the same time were important in differentiating among the groups.

The number and definitions of the personal qualities changed somewhat as a result of the field conferences. For example, the raters told us that they would find it difficult to differentiate between the traits Motivation and Initiative. Therefore, on the final form we combined the two into a single definition under the trait Initiative. An invariable request from the field settings was to include the trait "Military Bearing" in the final list of attributes to be rated. Likewise, we discovered that a number of the performance of duties items taken from the job inventory with good statistical properties were not particularly meaningful to raters in the field and were, therefore, deleted.

One thing that I think was unique about our approach is that we took these suggestions seriously. I know that there is a lot in the advice-giving literature in performance appraisal to suggest that you should avoid using personal traits on evaluation forms, particularly traits such as Military Bearing. However, one of our prime interests was in the attitudes of raters in the field. We felt that before you can expect raters to give you good data you must have an instrument that has credibility and is acceptable and meaningful to the people who are going to use them. (Moreover, on a more technical level, the suggestion by Kavanaugh that the evidence is not yet in on the superiority of behavior ratings versus trait ratings is due some consideration).

5. Experimental Test Period

The final aspect of this project that I think is important to relate to you is the commitment on the part of the Coast Guard for an experimental tryout of the proposed forms before they are implemented system-wide. The purpose of the field tryout will be to investigate the psychometric characteristics and the psychological reactions of the proposed system under "live" conditions. We will try to find out

whether or not the new system with seven forms can improve upon the existing system that I described at the beginning of this talk. As part of the current project, then, we have proposed an experimental design whereby the various aspects of the proposed system can be tested and evaluated.

Conclusions. Before I leave you this morning I'd like to comment on how we think this new appraisal system will solve some of the ills of the existing system that I outlined 45 minutes ago. First of all, we hope that grade effects in the marking system will be reduced simply by the fact that high ranking NCO's will now be evaluated on a separate form than lower ranking NCO's. We think this will make it more palatable for raters in the field to rate an E7 or an E8 as average, in comparison with other E7's or E8's in the Coast Guard. Secondly, we think that the tendency of raters to rate leniently will be reduced on a form that dramatizes the fact that 90 percent of the personnel should be rated in the middle three blocks on the form. If the feedback system works and raters begin rating average performance in the average block, we should get the kind of discrimination in the system that is needed in order to identify only the outstanding candidates for promotion. With respect to redundancy of information, we don't kid ourselves about the fact that we will get high correlations among the many different items on the form. However, by breaking down complex concepts such as Leadership and Proficiency we offer the rater the possibility to rate differentially within a single ratee, rather than to rate on the basis of an overall global evaluation. All these ideas, of course, will be tested in the experimental phase.

In summary, we have developed a promotion system for enlisted personnel that contains a rater feedback and monitoring component for a collection of seven separate evaluation forms. The unique aspects of this project that we have talked about this morning are 1) the philosophy regarding the most important determinant (rater motivation) of effective performance appraisal data, 2) the worker-oriented inventory that we developed, 3) the statistical methodology that we used to simultaneously analyze the three modes of the worker-oriented job analysis data, 4) the emphasis we have had on the importance of user acceptance, and 5) the field experimental tryout of the proposed system.

Table 1

Actual Contributions of Factors for Advancement in Rate During March 1976 E3-E8 Servicewide Competition for Advancement*

Factor	Percent Contribution Intended	Percent Contribution Actual
Examination Score	44	40
Performance Evaluations	28	15
Time in Service	11	38
Time in Paygrade	11	6
Medals and Awards	6	1

*Taken from J. F. Stumpff and R. D. Chavalier, *An Analysis and Proposal for Revision of the Coast Guard Enlisted Performance Evaluation System.* Thesis submitted to the Naval Post Graduate School, Monterey, California, December, 1976.

Table 2

Varimax Rotated Eigenvectors for the Two-Dimensional Approximation of the Grade Mode Variance

	Vectors	
Grade	I	II
E-4	.74	-.11
E-5	.59	.06
E-6	.30	.35
E-7	.01	.64
E8/E9	-.09	.67

Table 3

Response Categories on the Final Rating Form

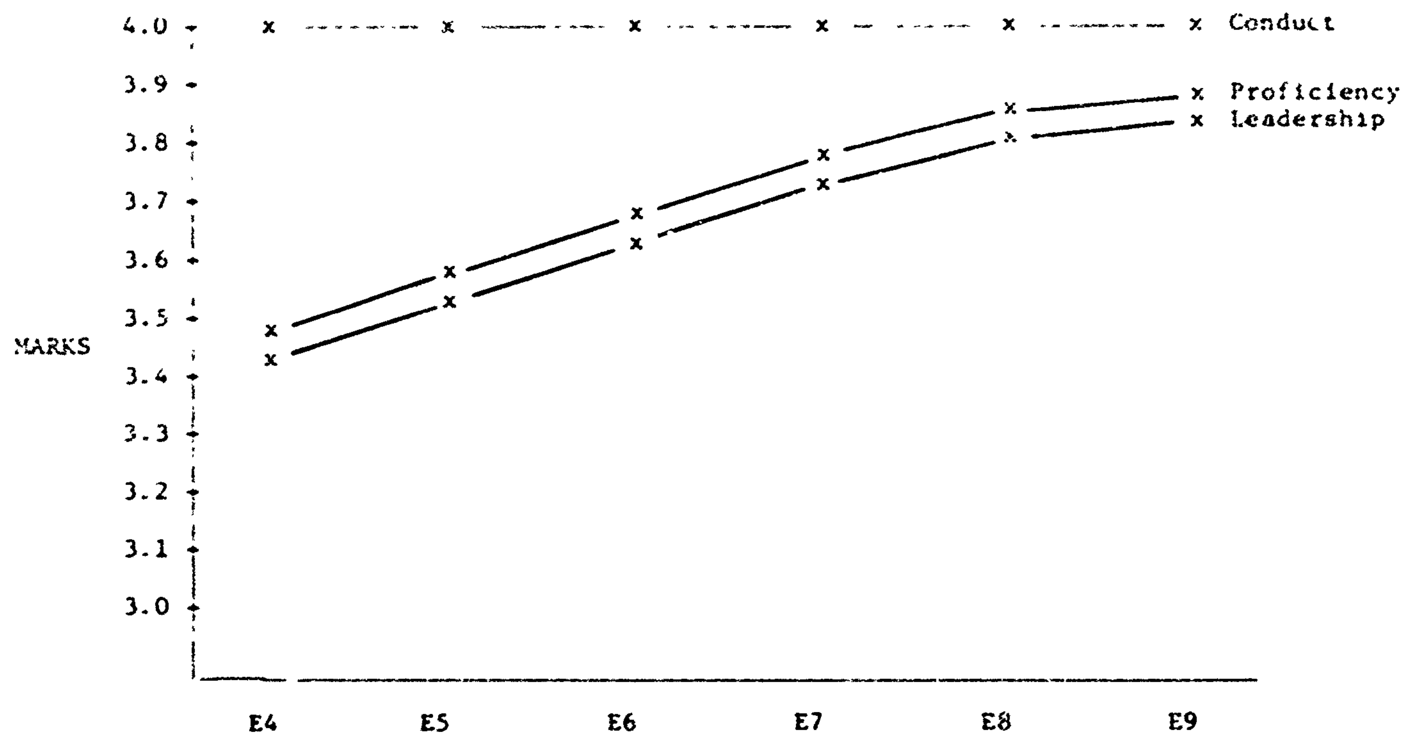
RESPONSE CATEGORIES

Not Observed	Unsatisfactory	Good			Outstanding
	5%	10%	70%	10%	5%

Figure 1

Analysis of Coast Guard-wide Sample of Performance
Marks for 31 December 1975*

N = 2230



*Taken from J. F. Stumpff and R. D. Chevalier, *An Analysis and Proposal for Revision of the Coast Guard Enlisted Performance Evaluation System*. Thesis submitted to the Naval Post Graduate School, Monterey, California, December, 1976.

Notes on the Feasibility of Predicting Fighter Pilot Effectiveness[1]

E. W. Youngling, S. H. Levine,
J. B. Mocharnuk and L. M. Weston
presented by J. B. Mocharnuk

McDonnell Douglas Astronautics Company - St. Louis
P.O. Box 516, St. Louis, Missouri 63166

ABSTRACT

History has demonstrated that there is a pressing need for improved selection and training of fighter pilots. In World War II, only one of twenty pilots became an ace. The U.S. Air Force kill ratio in Southeast Asia was approximately 2.5 to 1. In contrast, the Israelis claim to have a kill ratio of 60 to 1.

Our feasibility study has focused upon enhancing our record of air-to-air combat kills through more stringent and comprehensive selection procedures. Evidence is presented which demonstrates that a program can be developed to select pilots who will be effective in air-to-air combat.

From reviews of U.S. and foreign selection research dating from World War II to the present and an assessment of pilot opinion from hundreds of aces, 45 factors were identified as potential predictors of fighter pilot combat effectiveness. Of these 45 factors, only 10 are adequately evaluated within current military selection programs upon entrance into pilot training. Assessment of the remaining 35 untapped factors is within our technological reach. In fact, many of these factors can be assessed by tests which are presently available.

We developed an Air Combat Effectiveness Study (ACES) program which would establish selection test measures for virtually all of the factors identified as underlying fighter pilot combat effectiveness. As part of the ACES program, selection test measures would be validated against performance in air combat maneuvering ranges, thereby providing a method for selecting fighter pilots during peacatime. We have emphasized selection for success in the operational environment rather than success in training.

Armed with these selection test scores and an effectively executed validation program, researchers should, for the first time in history, be able to specify a definitive profile of the ace fighter pilot.

---

[1]This paper is based upon ARPA (Defense Advanced Research Projects Agency) Contract No. MDA-9-3-76-C-0169, "Feasibility Study to Predict Combat Effectiveness for Selected Military Roles: Fighter Pilot Effectiveness" by E. W. Youngling, S. H. Levine, J. B. Mocharnuk, and L. M. Weston, dated 29 April 1977, MDC Report E1634.

Notes on the Feasibility of Predicting Fighter Pilot Effectiveness

Fighter pilot combat effectiveness was selected for investigation for several reasons. Top level DoD concern with combat effectiveness is always present and the Defense Advanced Research Projects Agency was interested in the importance of manning high cost weapons systems. In those systems, the cost of developing and implementing proper selection and training programs should be quite small relative to the total weapons system cost. Because the McDonnell Douglas Corporation is intimately concerned with fighter aircraft systems effectiveness, the fighter pilot role was of special interest to us. Also, recently developed combat maneuvering ranges were considered, with simulators, as potential tools for use in selection as well as training. Complementing these interests, the history of air combat has demonstrated that a need for improved pilot selection and training exists.

In World War II we sent many pilots to war; some survived their early combat engagements, became skilled at their craft, and went on to become aces - but they were the exception. According to official Eighth Air Force records of approximately 5000 fighter pilots who flew against the Germans during 1943 - 1945, relatively few became aces. Only 261 (about 5.2 percent) achieved this goal. However, this small group of men were responsible for 40 percent of the total 5284.5 German planes destroyed by the Eighth Air Force fighter pilots during that period; thus, 5 percent claimed 40 percent of the kills (Eighth Air Force, 1945).

In the Korean War, once again it was found that a small percentage of Air Force pilots were obtaining most of the kills. Here again, the largest group of pilots recorded no kills (53.5 percent) while a small group of 38 pilots (4.8 percent) became aces. Importantly, each of the F-86 pilots had at least 25 counterair missions and, therefore, presumably, a fair chance for a kill (Torrance, Rush, Kohn, and Doughty, 1957). Clearly, the aces, a small group, make an overwhelming contribution to air-to-air kill records and air supremacy. We must find a way to augment our record of air-to-air kills, especially since we will probably be able to field only a relatively small number of fighter pilots in future wars.

Fighter pilots have recommended that the way to improve air-to-air performance is to select a man according to more rigid standards, give him specialized training, and keep him in the cockpit. We feel that it is prudent to seriously consider the recommendations of the fighter pilot community.

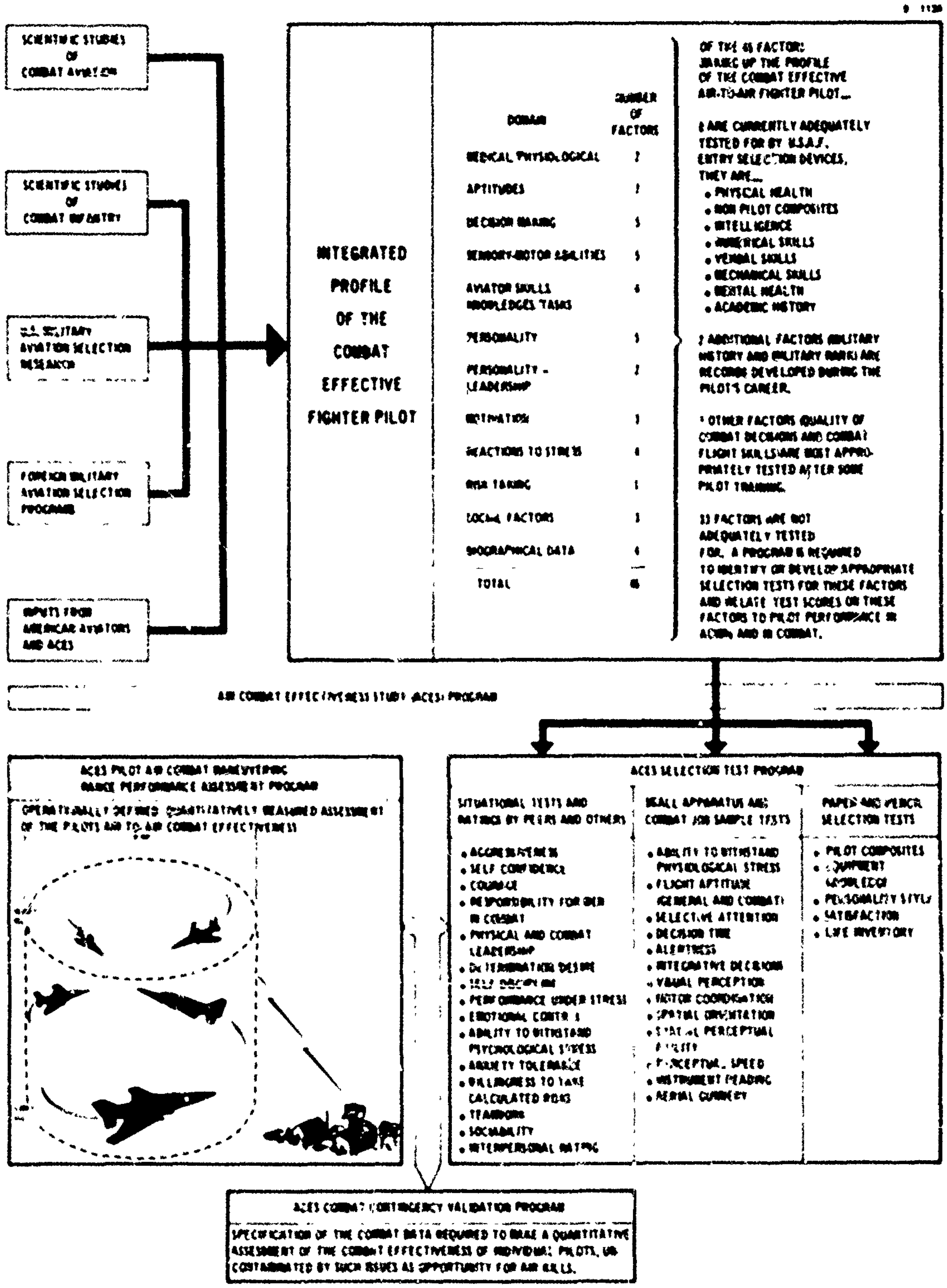


FIGURE 1 AIR-TO-AIR COMBAT EFFECTIVENESS FEASIBILITY STUDY

Much of the present study is concerned with the issue of finding a better way to select men for the fighter pilot role. We have emphasized air-to-air combat. The development of a selection program of this nature requires several discrete operations - a job analysis of the fighter pilot task, the generation of testable trait hypotheses, the development of predictor variables, validation, and cross validation. In this study, an extensive and comprehensive job analysis was performed, hypotheses were established, and programs were outlined for execution of the remaining operations. We conclude that a program can be instituted which could select men at entry into the military who would prove to be effective air-to-air pilots. For validation purposes, these pilots can demonstrate their effectiveness, during peacetime, in air combat maneuvering range engagements. High quality intermediate criteria can be developed in this type of facility. Furthermore, we believe it is possible to select pilots who have the motivational characteristics of those who will fight effectively in actual combat.

Figure 1 contains a portrayal of the classes of information evaluated, the resultant profile of the combat effective, and the ACES program. The principal work elements of this study focused on the identification of those critical characteristics and skills which are thought to characterize the combat effective pilot. In all, five major sources of information were used to generate the integrated profile of the combat effective pilot shown in Figure 1. This profile is based on our comprehensive and systematic review of World War II, Korean, and Southeast Asian conflict information and deals with the scientific studies of combat aviation and combat infantry for those wars. The U.S. Military Aviation Selection research from World War II to the present time was reviewed. This literature, which focuses upon characteristics and critical skills which predict success in flight training, has been used as a source of hypothesis generation because, by implication, these factors should be related to combat effectiveness. We also reviewed the German and Japanese World War II aviation selection research programs as well as the current Israeli program. Our final source of information came from 373 questionnaires which were returned to us by fighter pilots. Significantly, 280 of these returns were either from ace aviators or aviators with MIG kills in Southeast Asia. The fighter pilot organizations were surveyed and their response rates are reflected in Figure 2.

Using the data and inputs from the five sources of information, we generated the integrated profile of the combat effective fighter pilot which is shown in the accompanying figure. In all, some 45 factors distributed among 12 major domains (see Figure 1) can be reasonably hypothesized to be of predictive value in identifying the combat effective air-to-air fighter pilot. As an example, of the 45 factors which can legitimately be supposed to underlie air-to-air fighter pilot combat effectiveness, 35 of them are not adequately tested for by the U.S. Air

PILOT ORGANIZATION	NO. SENT	NO. BACK	%
AMERICAN FIGHTER ACES ASSOCIATION	524	257	49
NAVY MIG KILLERS ASSOCIATION	42	22	52
RED RIVER VALLEY FIGHTER PILOTS ASSOCIATION	98	51	51
NATIONAL GUARD PILOTS	40	13	33
ACEVAL – AIMVAL NAVY PILOTS	12	12	100
ACEVAL – AIMVAL AIR FORCE PILOTS	6	6	100
AGGRESSOR SQUADRON AT NELLIS	12	12	100
TOTAL	736	373	51

FIGURE 2 PILOT ORGANIZATION AND RESPONSES

Force in their entry selection program. Since the methodology either exists or is within our technological grasp for testing the bulk of these hypothesized predictor variables, we believe that a prima facie case has been made for the overall feasibility of such a research program.

In the lower right hand portion of Figure 1 is an outline of the ACES selection test program which would test for the 33 factors listed within the figure. The hypothesized predictor variables are grouped according to the class of testing device which emerged as most appropriate in our analysis of cost-effectiveness and practicality. We conceptualize a selection test program as having similarities with the current Israeli program, although probably more comprehensive. Such a selection test program could best be conducted at a single site, and candidates would be tested during an estimated seven to ten day period.

The next step in the ACES program, after the implementation of the combat effective pilot selection test battery, is the peacetime validation phase (air combat maneuvering range performance assessment program). The use of success in pilot training as a criterion for pilot selection

has proven unsatisfactory as a means of identifying the combat effective pilot. While successfully completing pilot training is a necessary condition for becoming a pilot, it is not a sufficient condition for becoming combat effective. A performance criterion having a stronger relationship to combat success is needed to properly validate the test battery. Evidence that tests which predict success in training also predict success in the operational environment has not been found. Figure 3 shows correlations between three predictors and two criteria, passing advanced flight training and the number of combat kills. Although the data are from

		CRITERIA	
		PASS-FAIL	COMBAT KILLS
PREDICTORS	PILOT STANINE	.6 .57	.00 —.12/.14
	DIAL & TABLE READING	.40	—.17/.13
	SPATIAL ORIENTATION	.34/.38	.08/.18

FIGURE 3 TRAINING CRITERIA VERSUS OPERATIONAL CRITERIA

different studies, upon inspection they suggest that we, at least, carefully evaluate the utility of using only training criteria for validating our selection instruments.

Once this phase of the peacetime program is completed, the pilot's combat effectiveness scores would be correlated with the ACES selection test program scores, and the determination made concerning the predictive power of the ACES selection test program. If, upon analysis, adequate correlations exist between some of the candidate pilot's selection test scores and his performance during dissimilar air combat testing, then there will exist, during peacetime, a way of selecting pilots who will perform adequately in a necessary condition for combat success. Combined with appropriately small selection ratios, the probability of selecting combat effective pilots will be greatly enhanced.

Finally, we have prepared an ACES combat contingency validation plan, for ultimately, should the situation eventuate, one could relate the pilot's scores in both the ACES selection test program and the ACES pilot air combat maneuvering range performance assessment program with performance in actual combat. The combat contingency validation plan includes

a specification of the combat data required to make a quantitative assessment of the combat effectiveness of individual pilots, uncontaminated by such issues as opportunity. We have used the Strawbridge and Kahn (1955) study of combat effectiveness in the air war in Korea as a model for developing the data requirements for a scientific and rigorous combat data collection program. Critical combat data requirements identified by those authors include missions, sightings as leaders, firings, and weighted kills.

While the overall ACES program is ambitious, it is, in our opinion, quite feasible and is potentially a very high payoff program. The nature and size of the effort required to accomplish this job is such that it will clearly require high level endorsement and sponsorship.

The ACES combat contingency validation program which we have sketched out here is clearly provisional. However, it does supply a usable departure point for a more carefully contrived plan. Air Force efforts, both in World War II and, particularly the Korean conflict, were well conceived and executed. Indeed, they form a large basis of what we can say factually about the factors contributing to air combat effectiveness. However, the researchers, through no real fault of their own, had only very limited relevant information on the pilots prior to their entry into combat.

If the ACES program recommended in this report is implemented, research scientists will be armed with selection test scores for virtually all of the dimensions presumed to underlie fighter pilot combat effectiveness. Armed with these test scores and an effectively executed combat data acquisition validation program, the military aviation research community should be able to select those persons who are most likely to be combat successful fighter pilots.

References

Final report of assessed fighter claims against enemy aircraft, August 1942, April 1945, Eighth Air Force, published September 1945.

Strawbridge, D. and Kahn, N. Fighter pilot performance in Korea. (AIWR Report No. 55-10). Chicago, Illinois: Institute for Air Weapons Research, November 1955.

Torrance, E. P., Rush, C. H., Jr., Kohn, H. B., and Doughty, J. M. Factors in fighter-interceptor pilot combat effectiveness (AFPTRC-TR-57-11). Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, November 1957. (NTIS No. AD-146 407)

Initial Development of the Organizational Assessment Package

William H. Hendrix

Air Force Human Resources Laboratory

Brooks Air Force Base, Texas

INTRODUCTION

Organizational effectiveness is an area of vital concern to the Air Force. It is an area that can not only result in financial loss but loss of human resources. Most Air Force personnel can recall personal experiences where they have witnessed organizations which were obviously inefficient financially as well as having morale and productivity problems.

The problem addressed in this paper is how does one effectively model organizational effectiveness and, in turn, measure it. Toward that end, a Three Component Organizational Effectiveness Model is presented and the data collection instrument package which is based on the model is described. The instrument package is entitled the Organizational Assessment Package (OAP).

BACKGROUND

The Organizational Assessment Package (OAP) is being developed for use by the Air Force's Leadership and Management Development Center (LMDC) Maxwell AFB, Alabama. The objectives of LMDC include: (a) providing consultative services to Air Force commanders, (b) provide Leadership and Management training to Air Force personnel in their work environment, and (c) perform research in support of (a) and (b) above. The consultative

role involves organizational problem area identification and recommendations for reducing or eliminating problems identified. The OAP is being designed to meet LMDC's objectives. First, the OAP will provide a means of identifying existing strengths and weaknesses within organizations. Second, research results can be fed back into their Professional Military Education, other leadership and management training courses, Air Staff, and functional Offices of Primary Responsibility (OPR's). Lastly, the OAP data base established can be used for research to strengthen the overall Air Force.

THREE COMPONENT ORGANIZATIONAL EFFECTIVENESS MODEL

The Three Component Organizational Effectiveness Model[1] (c.f. Figure 1) was primarily reported by Hendrix (1976), and considered Organizational Effectiveness (E) to be a function of: the criterion selected (c); the managerial style employed (m); and the situational environment (s), which includes the manager's subordinates, peers, and other personnel in the environment. That is: $E = f(c,m,s)$.

ORGANIZATIONAL ASSESSMENT PACKAGE

The Organizational Assessment Package (OAP) is designed to measure the basic components of the Three Component Organizational Effectiveness Model. As can be noted in Figure 1, the Supervisory Job Inventory (SJI)

[1] In Hendrix (1976) the model was initially entitled the Three Component Leadership Effectiveness Model and has since been expanded to focus on the entire organization.

is designed to measure managerial style (m), while the situational environment (s) is measured by two sections of the OAP, the Background Information section and the Organizational Job Inventory (OJI). The criteria are satisfaction, organizational climate, and perceived productivity. These are measured by the sections entitled; Job Satisfaction Questionnaire (JSQ), Organizational Climate Inventory (OCI), and Perceived Productivity Index (PPI). Hard data when available will be collected separately and merged with the OAP data base.

OAP FACTORS

Items within each of the OAP sections have been written to measure certain factors. The Background Information section contains biographical information items and items associated with factors in the situational environment. The factors in the situational environment which the items attempt to measure include: (a) organizational level of work group, (b) work group type, (c) work group size, (d) group member maturity, (e) Organization's geographic region, (f) extent to which work group meetings are used to establish goals, (g) extent of communication between work group members, and (h) stability of work hours. In addition, the situational environment is in part measured by the Organizational Job Inventory (OJI). The factors included in the OJI are based, in the main, on the job enrichment model proposed by Hackman, Oldham, Jansen, and Purdy (1975). They proposed five basic factors which they called Core Job Dimensions, Those were: (a) skill variety, (b) task identity, (c) task significance,

(d) autonomy, and (e) feedback from the job. These factors are to be measured by the OJI plus one additional work related factor which is labeled Work Interference. This factor deals with the extent and adequacy of: (a) additional duties, (b) equipment and supplies, and (c) provided work space.

In the criterion area, organizational climate is measured by the Organizational Climate Inventory which includes the factors of: (a) Communications, (b) general organizational conditions, (c) employee concern, (d) employee commitment, (e) decision making, and (f) recognition.

Another criterion area is that of job satisfaction which is measured by the Job Satisfaction Questionnaire (JSQ). This questionnaire contains 30 items which are descriptions of 30 factors out of 35 factors isolated by Gould (1975) in an unpublished study. The methodology and items used to isolate the factors can be found in Tuttle, Gould and Hazel (1975). The 30 factors are listed in Table 1.

The last criterion is perceived productivity and is measured by 4 items contained within the Perceived Productivity Index section. The items measure perceived productivity in terms of the work group's: (a) quantity of work output, (b) quality of work output, (c) performance when high priority work arises, and (d) whether flow of work to or from the work group is impaired.

The Supervisory Job Inventory (SJI) consists of 81 items relating to supervisory behavior. Once an adequate sample has been obtained these items will be factor analyzed and the resulting factors will be used to depict differing managerial behaviors.

PROGRESS

A small scale study (n = 144) was conducted at Lackland Air Force Base

during May 1977. One purpose of the study was to collect critique information on the OAP in order to improve it. In addition, the data served to provide an initial base line in terms of means and standard deviation for each item on the OAP. An intercorrelation matrix consisting of the OAP item variables plus a series of compound variables generated from the original variables, was used to: (a) delete items which did not intercorrelate well with the stated factors, and (b) establish simple correlational relationships between variables in the situational environment and managerial area with criteria items. The OAP previously described is the result of revisions based on data collected from the Lackland study. The major modification was the deletion of the Job Diagnostic Survey (JDS) (Hackman, et al, 1975) from the instrument package, with the OJI being used instead to establish the job enrichment variable values. The reason for deleting the JDS instead of the OJI was to reduce the total pages in the OAP (i.e. the JDS is approximately 7 pages and the OJI is 2 pages) and to have the format of the instrument the same as that of the other instruments within the OAP. The JDS is an excellent instrument and if the OJI indicates a job enrichment problem exists within an organization, then a more thorough examination could be accomplished using the JDS. Table 2 lists the intercorrelations between the job enrichment factors on the JDS with their counterpart on the OJI. Table 3 presents the intercorrelation of selected criterion items with the situational variables of: (a) a total score across items on the OJI (OJI Total), (b) the Motivation Potential Score (MPS) and Growth Need Score (GNS) as defined by Hackman et al, (1975), (c) the Need for Enrichment Index (NEI) which is derived for the OJI and is the total score of all items indicating a need for enrichment, and (d) the Job Motivation Index (JMI) which is computed with the same formula as the GNS.

DISCUSSION

The OAP is designed to provide indicators of the manager's behavior, the situational environment and criteria of organizational effectiveness. Should problem areas be identified then a more detailed investigation will be performed by on site consultation teams. Once validated the OAP should provide a means for: (a) identifying organizational strengths and weaknesses, (b) establishing appropriate managerial behavior in different situations with different criteria of success, and (c) identifying and resolving functional, career field, or systematic Air Force problems.

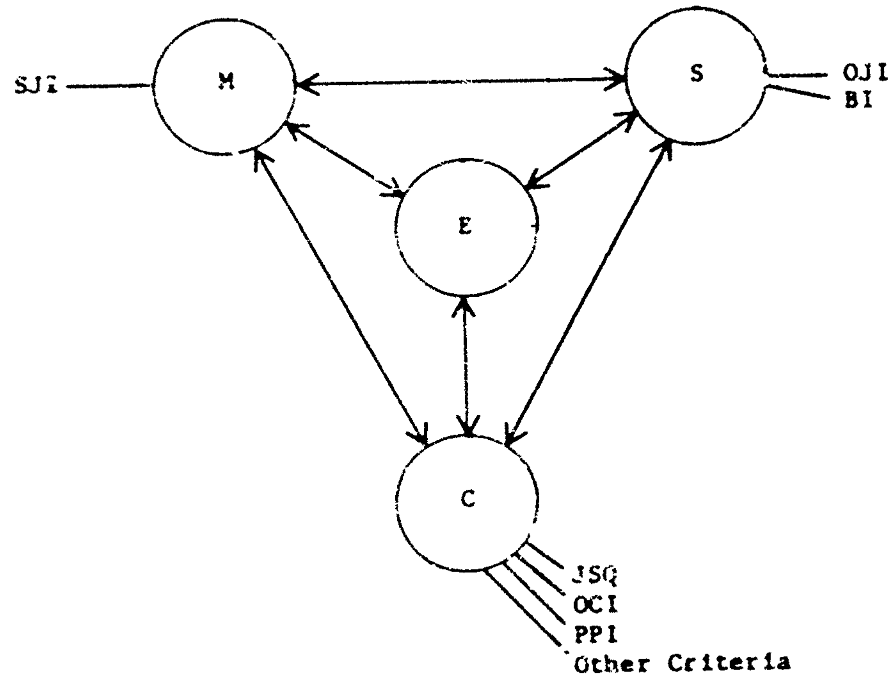
REFERENCES

Gould, R. B. Dimensions of job satisfaction: Development of the Air Force Occupational Attitude Inventory. Unpublished AFHRL research project 1975.

Hackman, J. R., Oldham, G., Janson, R., & Purdy, K. A new strategy for job enrichment. California Management Review, 1975, 17(4), 57-71.

Hendrix, W. H. Contingency approaches to leadership: A review and synthesis. AFHRL-TR-76-17. Lackland Air Force Base, Texas: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1976.

Tuttle, T.C., Gould, R. B., & Hazel, J. T. Dimensions of job satisfaction: Initial Development of the Air Force Occupational Attitude Inventory. AFHRL-TR-75-1. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1975.



Legend:

M = Management style
S = Situational environment
C = Criterion
E = Effectiveness
SJI = Supervisory Job Inventory
JDS = Job Diagnostic Survey
OJI = Organizational Job Inventory
BI = Background Information form
JSQ = Job Satisfaction Questionnaire
OCI = Organizational Climate Inventory
PPI = Perceived Productivity Index

Figure 1. Three component organizational effectiveness model

TABLE 1

JOB SATISFACTION QUESTIONNAIRE FACTOR ITEMS

Additional Duties	Temperature of Work Environment
Equipment and Supplies	Leave Policies
Information on Policies and Procedures	Work Itself
Feeling of Helpfulness	Work Schedule
Control of Others (Non-Supervisory)	Job Security
Characteristics of the Local Area	Safety Programs
Work Space	Travel
Social Contact (Other than Co-workers)	Acquired Valuable Skills
Co-Worker Relationships	Base Facilities
Family Attitude Toward Job	Base Housing and Eating Facilities
Independence in Work Procedures	
Job-Associated Training	
Job Hazards	
Moral Acceptability of Job	
Self-Improvement Opportunities	
Social Contact Opportunities	
Physical Activity	
Verbal and Written Communication	
Supervisor Responsibilities	
Temporary Duty (TDY) Costs and Conditions	

TABLE 2

CORRELATIONS OF JOB ENRICHMENT FACTORS ON THE JDS WITH THOSE ON THE OJI

FACTOR	CORRELATION COEFFICIENT
Skill Variety	.60
Task Identify	.60
Task Significance	.64
Autonomy	.44
Job Feedback	.59

TABLE 3

SITUATIONAL ENVIRONMENT VARIABLE PREDICTORS OF CRITERIA

	CRITERIA						
	Work Satisfaction	Perceived Productivity[a]				Climate[b]	
Situational Variable		1	2	3	4	1	2
OJI Total	.52	.26	.41	.30	-.12	.26	.22
MPS	.62	.21	.39	.32	-.21	.26	.22
GNS	.16	.21	.15	.16	.10	.14	.22
NEI	.23	.16	.19	.16	-.04	.18	.16
[illegible]	.56	.32	.44	.32	-.17	.30	.25

[a] Perceived Productivity

1 = Quantity of Work Output

2 = Quality of Work Output

3 = Performance when high priority work arises

4 = Efficiency in work flow from and to work group.

[b] Climate

1 = you are proud of organization

2 = you feel responsible for your organization

QUALITY OF LIFE IN THE U. S. AIR FORCE:

1977 vs. 1975

Charles W. McNichols
T. Roger Manley
Michael J. Stahl

Air Force Institute of Technology

In 1975 the authors reported on the development and initial application of a nine factor model describing the Quality of Air Force Life (QOAFL).[1] Initially, the model was constructed to provide a theoretical framework for the Air Force Management Improvement Group's (AFMIG's) surveys of Air Force military members, civilian employees, spouses of military members, and base commanders. It has since been used as the unifying theme for an Air Force wide survey of all commanders and in a second survey of Air Force military members, resulting in a total data base which includes over 50,000 responses. Data obtained from the most recent military survey effort, performed in the spring of 1977, offers an opportunity to examine the model for stability and to look at a longitudinal comparison of Quality of Air Force Life perceptions. These issues will be explored in this paper, along with some more detailed comparisons within the 1977 sample.

THE QOAFL MODEL

The nine dimensions hypothesized as encompassing the scope of the Quality of Air Force life are listed and defined in Figure 1. In application, each factor is presented to a survey respondent along with its definition and a pair of seven point scales to be used in reporting degree of importance and satisfaction associated with the factor. In the Air Force Quality of Life surveys, each of the nine factors was followed by a sequence of more detailed survey items related to the factor.

In the earlier (1975) surveys, an importance scale with seven responses ranging from "Low Importance" to "High Importance" was used. Most responses tended to cluster toward the "High Importance" end of this scale. While this result reinforced the authors' belief that factors of major importance had been chosen, the importance scale was not very useful for discrimination purposes. Therefore, in the 1977 survey of military personnel, the scale was changed to range from "Moderate Importance" to "Very High Importance" as shown in Figure 2. The rescaling resulted in the hoped for increased variance in the importance responses, but prohibits meaningful comparison with the

earlier reported importance levels. For this reason, analysis in the remainder of the paper is based only on the satisfaction responses for each of the nine dimensions.

Figure 1

QOAFL FACTORS

ECONOMIC STANDARD: Satisfaction of basic human needs such as food, shelter, clothing; the ability to maintain an acceptable standard of living.

ECONOMIC SECURITY: Guaranteed employment; retirement benefits; insurance; protection for self and family.

FREE TIME: Amount, use, and scheduling of free time alone or in voluntary associations with others; variety of activities engaged in.

WORK: Doing work that is personally meaningful and important; pride in your work; job satisfaction; recognition for my efforts and my accomplishments on the job.

LEADERSHIP/SUPERVISION: Has my interests and that of the Air Force at heart, keeps me informed; approachable and helpful rather than critical; good knowledge of the job.

EQUITY: Equal opportunity in the Air Force; a fair chance at promotion; an even break in my job/assignment selections.

PERSONAL GROWTH: To be able to develop individual capacities; education/training; making full use of my abilities; the chance to further my potential.

PERSONAL STANDING: To be treated with respect; prestige; dignity; reputation; status.

HEALTH: Physical and mental well-being of self and dependents; having illnesses and ailments detected, diagnosed, treated and cured; quality and quantity of health care services provided.

Figure 2

QUESTIONNAIRE EXAMPLE: FREE TIME

Please rate the degree of importance of free time to you and your degree of satisfaction with it based on the following description:

FREE TIME: Amount, use, and scheduling of free time alone, or in voluntary associations with others; variety of activities engaged in.

What degree of importance do you attach to the above?

```
     A.....B.....C.....D.....E.....F.....G
  Moderate           High         Very High
 Importance       Importance     Importance
```

To what degree are you satisfied with the FREE TIME aspects of your life?

```
     A.....B.....C.....D.....E.....F.....G
   Highly                         Highly
Dissatisfied        Neutral      Satisfied
```

ANALYSIS

In this section of the paper the nine factor QOAFL model will be examined from a factor stability standpoint, mean satisfaction levels reported in the 1975 and 1977 surveys will be compared, and some of the differences in QOAFL satisfaction levels for various subsets of the population responding to the 1977 survey of Air Force military personnel will be reported.

Model Stability

Although shifts in mean satisfaction levels are to be expected over time, the model will be most useful for longitudinal research purposes if the correlation structure of the nine factors is found to be relatively stable. To test the model for this factor stability, principal component analysis results for the 1975[2] and 1977 surveys were compared. Table 1 summarizes the two analyses. In both cases, two strong factors can be identified, explaining slightly over 50 percent of the total variance. The factor loadings after varimax rotation suggest similar factor interpretations derived from the two sets of data: a general measure of satisfaction with the work situation as the first factor, and a measure of satisfaction with economic aspects of life as an Air Force member as the second.

TABLE I

Principal Component Analysis: 1975 Military Survey N = 10,996

Factor	Eigenvalue	Cum % of Variance	Factor Loadings After Rotation: Dimension	Factor 1	Factor 2
1	3.60	40.0	Economic Standard	.16	.77
2	1.04	51.5	Economic Security	.10	.82
3	.81	60.6	Free Time	38	.41
4	.77	69.1	Work	.77	.10
5	.65	76.3	Leadership	.71	.14
6	.62	83.1	Equity	.54	.38
7	.56	89.4	Personal Growth	.73	.26
8	.51	95.0	Personal Standing	.73	.26
9	.45	100.0	Health	.25	.47

Principal Component Analysis: 1977 Military Survey N = 10,687

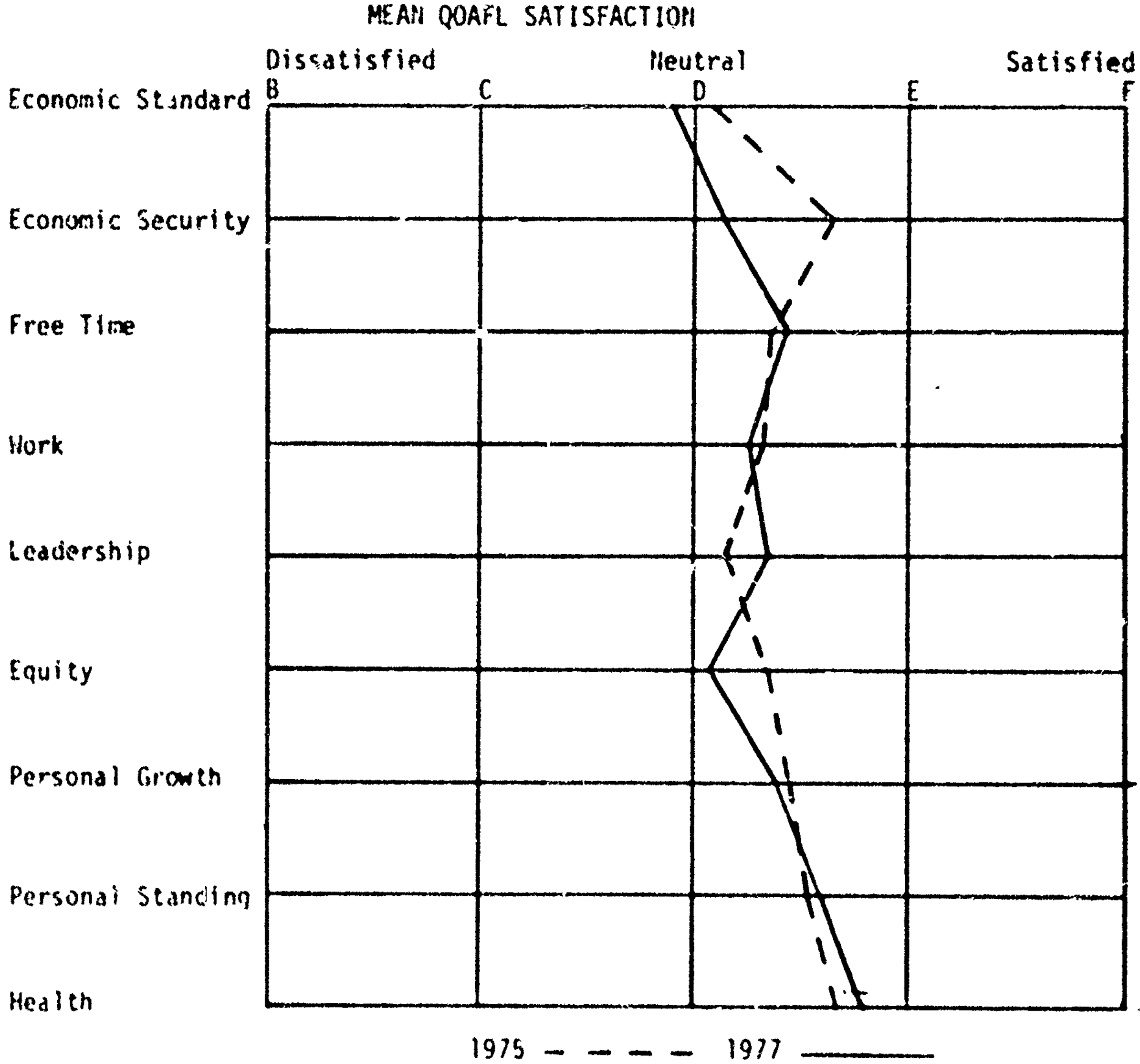
Factor	Eigenvalue	Cum % of Variance	Factor Loadings After Rotation: Dimension	Factor 1	Factor 2
1	3.50	38.9	Economic Standard	.15	.80
2	1.05	50.6	Economic Security	.12	.82
3	.83	59.8	Free Time	.50	.20
4	.80	68.7	Work	.72	.15
5	.66	76.1	Leadership	.70	.01
6	.64	83.2	Equity	.54	.38
7	.56	89.4	Personal Growth	.71	.28
8	.49	94.9	Personal Standing	.71	.25
9	.46	100.0	Health	.31	.45

Comparison of 1975 and 1977 Satisfaction Levels

Although the correlation structure of the QOAFL satisfaction scores did not change significantly between 1975 and 1977, there were some shifts in means for particular factors. In Figure 3 a profile diagram has been used to illustrate the direction and relative magnitude of these shifts. The diagram represents the mean scores for the overall sets of respondents to the two surveys for each of the nine factors. As can be seen from this diagram, the largest shifts in satisfaction levels occurred in the Economic Standard, Equity, and Economic Security dimensions, and were in the direction of higher dissatisfaction. All other shifts were small in magnitude, with slightly higher mean satisfaction reported with Leadership/Supervision, Personal Standing, Free Time, and Health aspects of respondents' lives, and slightly lower mean satisfaction reported with the Work, and Personal Growth factors.

Figure 3

Comparison of Mean QOAFL Satisfaction Levels Reported in 1975 and 1977



Differences Among Subsets of 1977 Respondents

In Figure 4 1977 mean satisfaction scores for officer and enlisted personnel have been plotted on the profile diagrams. Officer personnel report significantly higher satisfaction with Economic Standard, Work, Personal Growth and Standing than enlisted respondents, but mean satisfaction scores on all other factors are quite similar for the two groups.

As a final example of comparative QOAFL satisfaction levels, the mean scores reported in the 1977 survey by first term Air Force personnel have been plotted for each level of career intent reported in the same survey. The subset of the 1977 sample, in this case, represents all enlisted and non-rated officer respondents with less than four years service, and rated officers with less than six years service (because of the longer service obligation for flight training, rated personnel with under six years service were considered first termers).

FIGURE 4

Comparison of Mean QOAFL Satisfaction Levels for AF Officer and Enlisted Personnel, 1977

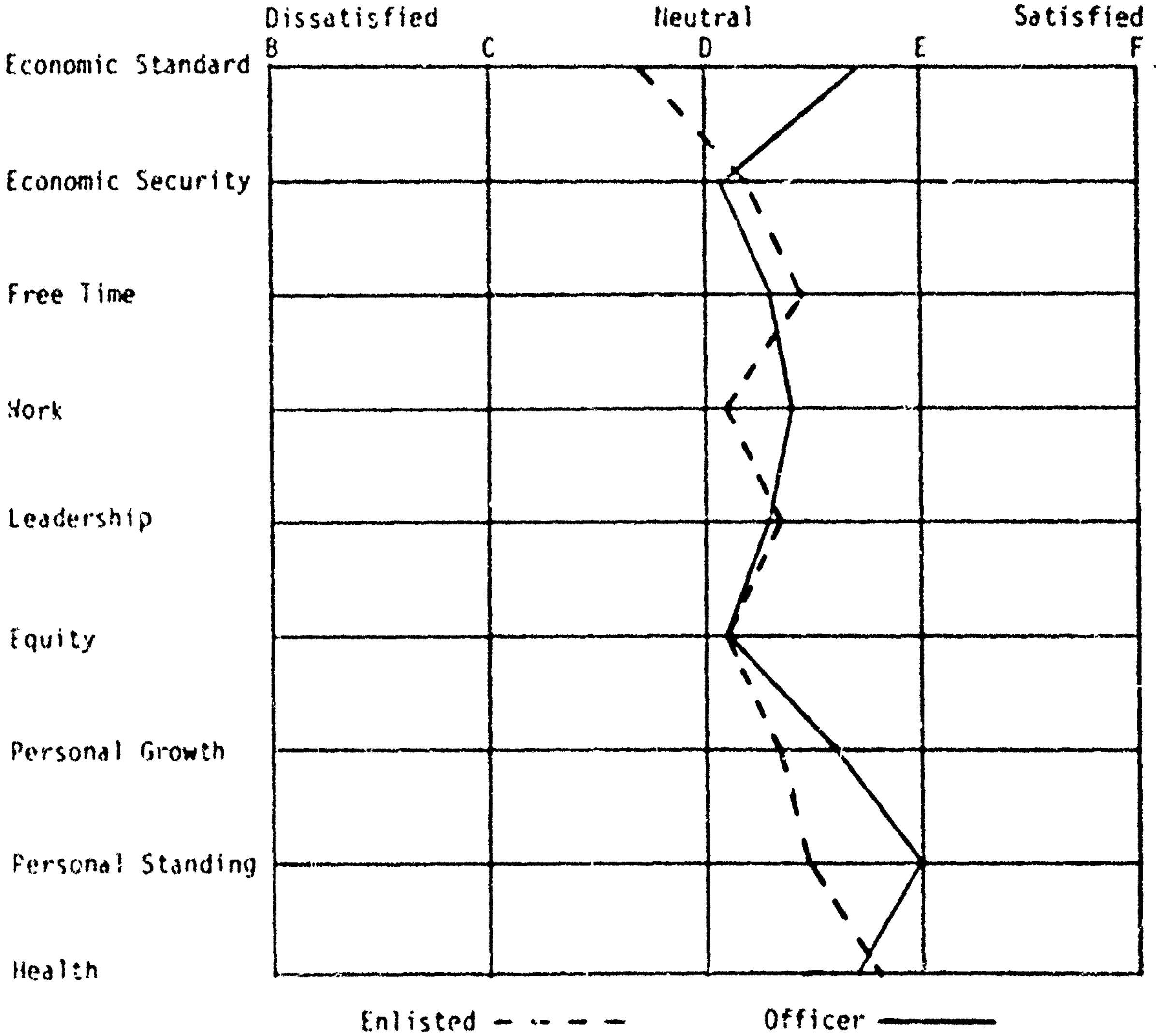
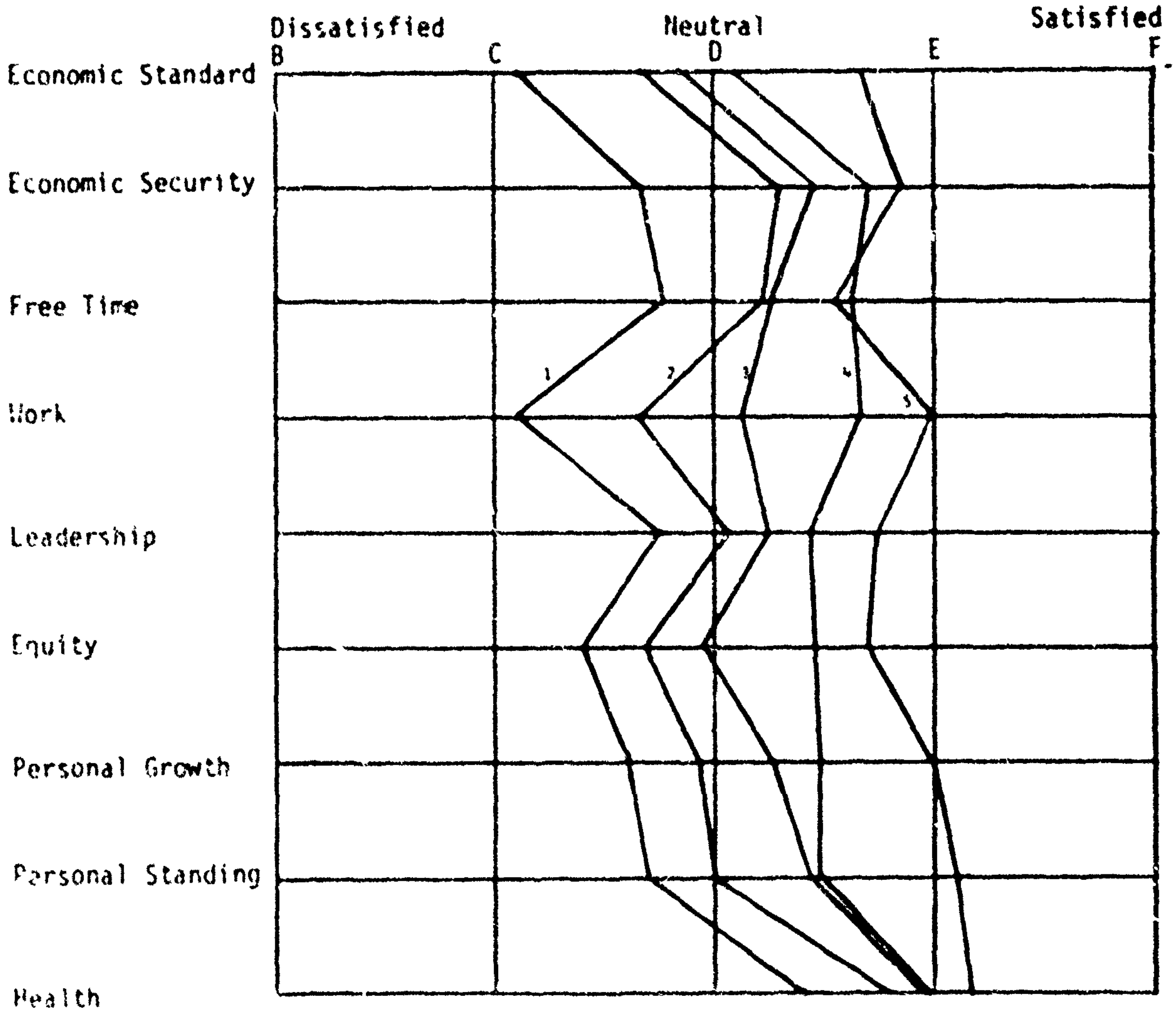


Figure 5 indicates that, while higher levels of career intent are associated with higher levels of satisfaction for all nine factors, Work and Economic Standard appear to be the most powerful discriminators of career intent level.

FIGURE 5

QOAFL Profiles for First Term AF Personnel by Reported Career Intent in 1977



[1]Definitely do not intend to make the AF a career
[2]Most likely will not make the Air Force a career
[3]Undecided
[4]Most likely will make the Air Force a career
[5]Definitely intend to make the AF a career

DISCUSSION

The nine factor Air Force Quality of Life model has now been used in six major Air Force surveys performed over a two year period. While there have been some changes in mean satisfaction levels for the nine factors during this interval, the basic correlation structure of the model appears virtually unchanged. Use of the 1975 satisfaction data as a baseline, and continued use of the model as a framework for Air Force opinion and attitude surveys seems justified.

NOTES

1. Manley, T. R., R. A. Gregory, and C. W. McNichols, Quality of Life in the U.S. Air Force, Proceedings, 1975 Military Testing Association Conference, Indianapolis, Indiana, Sept. 1975.

2. McNichols, C. W., T. R. Manley and R. A. Gregory, Measuring the Quality of Life of Air Force Personnel, Proceedings, 1976 Psychology in the Air Force Symposium, USAF Academy, Colorado, 8-10 April 1976.

Measuring The Quality of Navy Life

by

Richard J. Orend, Robert N. Gaines,
Kenneth W. Stroad and Marsha J. Michaels

INTRODUCTION

Ultimately, the military's interest in the quality of life reduces to two basic questions: (1) will improving the quality of life bring increased reenlistment rates and, by extension, greater enlistment interest, and (2) how is improvement in the quality of military life related to the on-the-job productivity of military personnel? If it can be shown that significant improvements will occur in these areas as a result of changes in the perceived quality of military life, then extensive research efforts will have been vindicated. If, however, the results of these efforts are simply nice to know information and "interesting" correlations, the resources spent on this research might be put to more fruitful uses. Of course, the eventual achievement of goals as ambitious as increasing reenlistment rates and productivity requires the cooperation of both researchers and policy makers, since the findings of any research efforts must be translated into concrete policies and implemented in real environments. Thus, researchers must operate within the constraints of feasible policies and policy makers must be willing to experiment and modify some traditional ideas and procedures if useful results are to be forthcoming.

Our purpose here is to examine efforts to develop the first stages of this process, namely, the measurement of the quality of military life. There are two distinct elements to this development, conceptual and methodological. Previous efforts to develop quality of life measures in the military have suffered because they generally ignored the conceptual aspects of the development process. The most important implications of this omission are the failure to treat all aspects of the quality of life which might be relevant to reenlistment decisions and productivity and the absence of a means to evaluate the lists which were developed. Essentally, there was no basis to judge, a priori, the inclusion of particular elements of life quality and there was no structure to serve as a heuristic by which additional variables or

dimensions could be evaluated. This led to instruments which excluded a large number of potentiality useful variables and to the measurement of what were presumably similar concepts with rather divergent indicators.

Another conceptual problem which has received insufficient attention is the dicision process by which perceptions of military life are transformed into decisions about behavior. Of particular importance there are questions about the relationship of job and non-job activities and the context in which decisions about reenlistment are made. That context includes the alternative courses of action open to individuals, the relative improtance of each of the factors in the quality of military life, experiences in the military, and the fulfillment of expectations about what military life would be like. Each of these factors can influence an individuals' evaluation of military life, i.e., its quality, and decisions about whether to remain in the military.

As is evident from the foregoing discussion the approach we follow is very broad and is intended to include all factors which may influence quality of life perceptions. This approach represents our initial attempt to identify a broad range of variables which may influence the behavior of military personnel and to examine interactions between perceptions of different aspects of military life, and between those perceptions and the context in which they are made. Our particular emphasis on all elements of the military life situation does not preclude narrow approaches which focus on one or a limited number of the factors which we feel are relevant to the descussion of the quality of military life.* In the following discussion an initial attempt on developing a general model will be described.

MEASURING THE QUALITY OF LIFE: A CONCEPTUAL FRAMEWORK

The lessons learned, both from examining the theoretical and methodological issues inherent in the previous research and from inspecting actual components of research instruments employed in

---

* Work by David Bowers, which focuses on the job related aspects of Navy life, is an example of the more restricted approach which has produced useful results.

these studies, will be applied in the following to the construction of a conceptual framework applicable to the measurement of the quality of Navy life. The process through which this framework will be fashioned involves: (1) establishing a theoretical structure which provides a rationale of life quality assessment; (2) identifying a set of life quality factors which adds substance to the theoretical structure; and (3) explaining how the resultant conceptual framework satisfies each requirement inherent in measuring quality of life.

A Theoretical Structure for Measuring Quality of Life: The theoretical structure offered here for the measurement of life quality has for its foundation the assertion that the quality of an individual's life is a positive function of the degree to which the individual's needs are satisfied. Thus, if nearly all of an individual's needs are being met, then his evaluation or expressed satisfaction with the quality of that life will be very high. If almost none of his needs are being met, then the evaluation of his life quality will be very low.

Based on the assertion above, the notion of quality of life here receives its primary structure from its analysis into several need categories. While a number of perhaps equally informative need taxonomies exist,* the most commonly accepted and frequently employed scheme of categorization is that proposed by Maslow.** This analysis will follow an approach adopted by several other quality of life studies by utilizing catergories which reflect only slight deviation from the pattern established by Maslow's need hierarchy.***

---

*On this point see Arnold Mitchell, "Life Ways and Life Styles" (Menlow Park, CA: Standord Research Institute, 1973), p. 5.

**See Abraham H. Maslow, "Motivation and Personality" (New York: Harper and Row, Publishers, Inc., 1954), pp. 80-98.

***Instances of studies which follow Maslow's categorization of needs include Angus Campbell, "Aspiration, Satisfaction, and Fulfillment," The Human Meaning of Social Change, Ed. Angus Campbell and Philip E. Converse (New York: Rusell Sage Foundation, 1972), 441-466, and Patricia A. Pecorella, Predictors of Race Discrimination in the Navy (Ann Arbor, Mich: Institute for Social Research, University of Michigan, 1975).

The four categories used will be termed: (1) safety and comfort; (2) belonging and love; (3) esteem; and (4) self-actualization. It is with respect to these categories, which serve as sub-scales of life quality, that overall quality of life will be measured.

Having received primary structure form an analysis of its conceptual contents into need categories, the notion of quality of life achieves secondary structure when these categories are themselves analyzed to reflect the logical distinction which exists between "quality of life" and "quality of work." This distinction is based on the premise that some different factors impinge on our lives in work and nonwork situations and insofar as this condition exists our evaluation of these life dimensions and the context in which they operate should be separately evaluated. In the military this distinction may be somewhat less prononced because of the overall control exercised on various elements of behavior, such as family separation and living and working on post often with the same supervisors.

The result of this secondary analysis is a conceptual matrix which permits assessment of both quality of life and quality of career with respect to each of the need categories. Table I provides a general representation of that matrix.

Factors in Quality of Life/Work: Furnished above has been a theoretical structure for the notion of life/work quality assessment. The objective now is to supply a set of factors which may be utilized as specific measures of quality of life/work. The factors may be generated by means of the following procedure. First, the component variables from each of the civilian and military related quality of life/work studies may be analyzed on the basis of their general content and logically associated into groups of similar variables. The crucial concepts common to groups of variables were then isolated and identified as preliminary life/work factors. Next, to this preliminary group was added another group of factors discovered in an initial analysis of the need categories furnished by the theoretical structure. The resultant factor set, which is composed of 39 elements, is illustrated in Table II along with corresponding variables from the studies treated above.

The quality of life/work factors set having thus been presented, an observation with respect to the exhaustiveness of this set is in order. In Table II, the factor set not only exhausts each of the variables utilized to assess quality of life/work of the military related

TABLE I

THE CONCEPTUAL FRAMEWORK

	LIFE	WORK
SAFETY AND COMFORT	Health and Medical Care Personal safety Living Essentials Local Environment Convenience	Income Secure Employment Retirement, Medical, and other Fringe Benefits Work Environment Job Convenience Sufficient Resources to Perform Job Organizational Climate Competence of Supervisor
BELONGING AND LOVE	Contribution to Community and Society Social Life and Relationships Relationships with Close Friends Relationships with Nuclear Family	Interpersonal Relationships in the Work Environment Work Related Friendships Family Disruption
ESTEEM	Self-Esteem Freedom of Choice and Expression Equality	Authority Responsibility Occupation Related Prestige Freedom to Decide How Work Should Be Done Participation in Decisions Affecting Own Future Meaningful Work
SELF ACTUALIZATION	Cognitive Development Affective Development Recreation Travel	Skill Development Utilization of Personal Skills Opportunity for Advancement Advancement on the Basis of Merit Interesting Work Creative Experience

TABLE II

FACTORS OF QUALITY OF LIFE/CAREER

Factor Set	Wilson, Flanagan, and Unlaner	Petty and Shell	Holz and Gitler	House, Livingston, and Swinburn	Survey Research Center
Health and Medical Care	Health and Personal Safety	Medical Plans & Fringe Benefits	Being provided with good Medical and Dental Care Facilities	Health Hazardous Substances	
Personal Safety		Personal Physical Safety		Safety	
Living Essentials		Owning a Home	Having Decent Housing and privacy in the Barracks Having good quality, sufficient quantity, and proper service of food.	Housing Accumulated Assest Privacy Essential Living Costs	
Local Environment				Ecosystem Land Use Climate Noise Water Pollution Air Pollution	
Convenience			Having Facilities Available on the Post that make Life Easier	Public Transportation Transportation Services	
Contribution to Community and Society	Social, Community, and Civic Activities	Involvement in Community Life			
		Opportunity to make a Lasting Contribution to Society			
Social Life and Relationships	Relations with Parents, Siblings, or other Relatives				
	Socializing				
Relations with Nuclear Family	Relations with Spouse (or girlfriend(s) boyfriend(s)).			Primary Social Relationships	
	Having and Raising Children				
Relations with Close Friends	Relations with Close Friends			Secondary Social Relationships	

TABLE II
(Continued)

Factor Set	Wilson, Flanagan, and Uhlaner	Petty and Sheal	Hols and Citler	House, Livingston, and Swinburn	Survey Research Center
Self-Esteem	Understanding and Appreciating Self		Being Treated like an individual and not like another number		
Equality			Being paid a fair salary equal to what Civilians make Getting Equal Treatment Regardless of Race	Equality	
Freedom of Choice and Expression	Philosophical and Ethical Values	Personal Freedom in Expression of Ideas	Being able to Cut one's Hair the Way one Wants Shortening the Length of a Tour and Letting one Choose the Location	Choices in Life	
			Being able to be what one Wants to be on one's own Time	Use of Free Time Leisure	
Cognitive Development	Intellectual Activities	Opportunity for further Civilian Schooling	Having Educational Opportunities and Post-Discharge Educational Benefits	Education	
		Opportunity to Develop into a Well-Rounded Individual, Opportunity to Realise Maximum Potential			
Affective Development	Aesthetic Activities			Culture	
	Spiritual Experiences or Beliefs			Cultural and Spiritual	
Travel		Opportunity to Travel			
Recreation	Active Recreational Activities; Passive Recreational Activities			Recreational Resources	

TABLE II
(Continued)

Factor Set	Wilson, Flanagan, and Uhlaner	Petty and Snell	Holz and Gitler	House, Livingston, and Swinburn	Survey Research Center
Income	Material Well-Being and Security for the Future	Pay		Income Income Distribution; Discretionary Income	The Pay is Good
Secure Employment		Steady Work		Employment Economic Security	The Job Security is Good
Retirement, Medical, and Other Fringe Benefits		Retirement Plan		Accumulated Assets	My Fringe Benefits are Good
Work Environment					Physical Surroundings are Pleasant
Job Convenience		Graphic Location of Job			The Hours are Good; Travel to and From Work is Convenient
Organizational Climate		Policies of Organization toward Employees	Getting Rid of Rules and Regulations that Don't Help Performance		I am free from conflicting Demands that other People Make of me
Sufficient Resources to Perform Job					I receive enough Help and Equipment to get the Job Done; I have enough information to get to get the Job Done
Competence of Supervisor		Technical Ability of Supervisor	Having Officers and Non Commissioned Officers that Know their Jobs		My Supervisor is Competent in Doing his Job
Family Disruption		Lack of Family Separation			
Interpersonal Relationships in the Work Environment		Good Interpersonal Relationships with Supervisors, Good Interpersonal Relationships with Peers, Good Interpersonal Relationships with Subordinates			I am give a lot of Chances to Make Friends; My Coworkers are Friendly and Helpful

TABLE II
(Continued)

Factor Set	Wilson, Flanagan, and Uhlaner	Petty and Shell	Holz and Gitler	House, Livingston, and Swinburn	Survey Research Center
Work Related Friendships					
Meaningful Work	Occupational Role (Job)		Making the Work Meaningful and Worthwhile and Eliminating the Busy Work		I can see the Results of my Work; The Problems I am asked to solve are hard enough
Responsibility		Amount of Personal Responsibility			My Responsibilities are clearly defined
Authority		Opportunity to be a leader; Opportunity to Control and Direct Others			I have enough Authority to do my job
Occupation Related Prestige		Highly Respected Job		Status	
Participating in Decisions about Own Future		Participating in Decisions about Own Future		Democratic Process	
Freedom to Decide How Work Should be Done		Freedom to do the Job the Best Way	Being able to do one's work without having to "hurry up and wait"		I am given a lot of Freedom to Decide How I do my Work
Skill Development	Developing Skills in Manual Areas	Chance for Training and Learning on Job		Personal Skills	I have an Opportunity to Develop my Special Abilities
Utilization of Personal Skills		Ability to Use Own Technical Skills			I am given a Chance to Do the Things I Do Best
Opportunity for Advancement		Chance for Advancement			
Advancement on the Basis of Merit		Fair Evaluation of Performance	Being able to Advance without having to "know the right people	Economic Opportunity	

TABLE II
(Continued)

Factor Set	Wilson, Flanagan, and Uhlaner	Petty and Shell	Hole and Gitler	House, Livingston, and Swinburn	Survey Research Center
Interesting Work		Interesting Work			The Work is Interesting
Creative Experience	Creativity	Producing Original Results or Products			

[1] Wilson, Sandra, Flanagan, John, and Uhlaner, J. E. Quality of Life as Perceived by 30 Year Old Army Veterans. Arlington, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences, 1975.

[2] Petty, M. M., and Shell, Timothy. The Use of Expectancy Theory in the Explanation of Turnover in ROTC. Journal of Vocational Behavior, 6 (1975).

[3] Hole, Robert F., and Gitler, George. Assessing the Quality of Life in the U.S. Army. Arlington, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences, 1974.

[4] House, Peter, Livingston, Robert, and Swinburn, Carol. Monitoring Mankind: The Search for Quality. Behavioral Science 20 (1975).

[5] Robinson, John, Athanasiou, Robert and Head, Kendra. Measures of Occupational Attitudes and Occupational Characteristics. Ann Arbor, Michigan: Institute for Social Research, 1969.

studies, but also includes 80 percent of the variables employed to measure quality of life and quality of career in the civilian related studies. In this way, the factor set displays a clear superiority of extension over the various sets of life/work quality variables used in the military related studies, and demonstrates a coverage of the variables critical to quality of life and quality of work measurement which is roughly equivalent to the more specialized civilian related studies. Second, despite the degree to which the factor set exhausts variables relevant to the assessment of quality of life/work, it must be considered a provisional set. This is because certain factors may be added or subtracted from the set based on the results of empirical investigation, conclusions derived from logical inspection of the theoretical structure, or specific research requirements.

MEASURING NAVY QUALITY of LIFE/WORK A RESEARCH DESIGN

General Approach

The foregoing analysis provides a basic model for the study of the quality of life/work in any context. In the proposed model we focus on the satisfaction of individuals with their military (Navy) lives, in both life and career situations. An analysis built on this framework can provide the basis for a relatively easy to administer general test for use with Navy personnel.

The focus of this discussion is both substantive and methodological. Substantively, we seek to specify some of the major problems confronting the Navy in terms of general satisfaction of personnel. Our concern is to first identify the general factors which comprise the total life space of Navy personnel, then to determine which of those factors is most closely associated with behavoral decisions, specifically the decision to reenlist.

Comparative Analysis

One of the most important methodological considerations in this research will be the use of comparison. That is, we want to analyze satisfaction not just with the Navy per se, but in comparison with what is expected in the civilian world, the standard against which individuals will be evaluating Navy life. Certain aspects of the Navy, e.g. pay, may displease everyone, but the relevance of a particular perception becomes important only when there is an alternative which is perceived as both better and available. Thus, we expect to be able to learn more about reenlistment decisions from a comparison of

Navy and civilian alternatives than from a Navy evaluation alone.

Other Contextual Factors

In a similar vein, each of the other contextual considerations mentioned previously is potentially important in the analysis of perceptions of the quality of Navy (military) life. For example, a difference in the perceived ability of the Navy to provide free choice in jobs vs. civilian choice is important only insofar as that freedom is significant to the individual. Another more popular example is the question of hair length. Most of the young men in the Navy feel that hair length regulations are restrictive, more restrictive than in civilian life. However, whether or not this perception is important in a reenlistment decision is at least partially a function of how important hair length is to the individual. We shall call this particular contextual consideration salience.

Another consideration is the set of expectations about Navy service enlistees brought with them. If I entered the Navy expecting to fly airplanes and ended up chipping paint, it seems likely that I would be greatly disatisfied with at least the work dimensions of my Navy career. While the discrepency may not be that large in most cases, there are undoubtedly many instances in which the reality of Navy life did not correspond with the expectations. At a minimum we would expect that such considerations would color evaluation of the Navy in the specific area where differences occur. They could influence Navy-civilian comparisons as well.

Still another part of the decision context is what actual experiences individuals had while they were in the Navy. By experience we mean in the institutional sense, such as rating, proportion of sea duty, and schooling, rather than the day-to-day interactions with peers and supervisors. The latter type of experience will be reflected in the specific variables evaluated by each individual and would not necessarily be associated with such general characteristics as rating. The former experiences are related to the constant impact of being at sea or working a particular type of job. While the previous context factors had to be measured and analyzed simultaneously with perceptions of quality of life variables, these experiences can be evaluated on a post-hoc basis by dividing respondents into groups which exhibit each of the relevant characteristics.

SUMMARY OF THE CONCEPTUAL MODEL

The foregoing discussion may be summarized as follows:

(1) Behavior of Navy personnel with regard to a reenlistment decision is a function of perceptions of Navy life modified by each individual's comparison of each variable to alternatives in civilian life and to the importance of that variable in their hierarchy of values.

(2) The variables which exhibit potential signifigance in these decisions may be identified through the use of a needs model which specifies the areas which are likely to be important to various groups of Navy personnel. Such a model helps to insure the comprehensiveness of the variable list and a systematic balanced approach.

(3) Beyond these basic considerations are such factors as expectations, and Navy experience, which may color the perceptions of individuals and thereby influence reenlistment decisions.

The usefulness of quality of life research will depend on our ability to account for each of these factors in a systematic way. By systematic we mean to evaluate decisions so that the impact of each of these factors can be identified and measured. From this base it will be possible to generate policy which reflects the reasons for negative evaluations of the Navy and the precise means to turn such evaluations (and presumably behavior) around.

"MEASUREMENT OF LEARNED BEHAVIORS IN COMPETENCY BASED LEADERSHIP TRAINING PROGRAMS"

A PRESENTATION

BEFORE THE

MILITARY TESTING ASSOCIATION

BY

DOROTHY VON K. PEPPER, ED. D.
WORTH SCANLAND, PH. D.

FOR MANY YEARS THE U. S. NAVY HAS BEEN ADDRESSING THE PROBLEM OF TRAINING LEADERS, AS HAVE MANY OTHERS IN AND OUT OF THE MILITARY SERVICES. THERE ARE THOSE WHO CONTEND THAT LEADERS ARE BORN, NOT MADE, AND THERE ARE OTHERS WHO TAKE THE POSITION THAT LEADERSHIP IS A DEFINABLE SKILL WHICH CAN BE IDENTIFIED AND TAUGHT, AND THAT THE RESULTANT BEHAVIOR CAN BE MEASURED. THE NAVY HAS BEEN IN BOTH CAMPS AT ONE TIME OR ANOTHER, BUT AT THE PRESENT TIME HAS ADOPTED THE POSITION THAT LEADERSHIP CAN BE DEFINED AS A LEARNED BEHAVIOR, SUBJECT TO IMPROVEMENT THROUGH TRAINING, AND THAT THESE LEARNED SKILLS ARE MEASUREABLE IF VIEWED AS COMPETENCIES. HAVING REACHED THAT CONCLUSION, THE NAVY,

AND SPECIFICALLY THE CHIEF OF NAVAL PERSONNEL, HAS COMMISSIONED MCBER AND COMPANY OF BOSTON, MASSACHUSETTS, TO DEVELOP A COMPETENCY BASED LEADERSHIP TRAINING CURRICULUM WHICH CAN BE DELIVERED TO BOTH ENLISTED AND COMMISSIONED PERSONNEL AT SEVERAL LEVELS OF SENIORITY, PRIMARILY AT SCHOOLS ASSOCIATED WITH ENTRY POINTS INTO THE NAVY OR INTO HIGHER LEVELS OF RESPONSIBILITY. IT IS THE PURPOSE OF THIS PAPER TO DESCRIBE THE METHOD BY WHICH THIS TRAINING PROGRAM IS BEING DEVELOPED, THE MEANS BY WHICH THE LEARNED SKILLS ARE TO BE MEASURED, AND THE CONTRAST BETWEEN THIS AND MORE TRADITIONAL APPROACHES TO TRAINING PROGRAM DEVELOPMENT.

IN A PAPER APPEARING IN THE JANUARY, 1973, ISSUE OF THE AMERICAN PSYCHOLOGIST, DR. DAVID C. MCCLELLAND OF HARVARD UNIVERSITY TOOK UMBRIDGE AT THE CONCEPT SO PREVALENT THEN AS WELL AS NOW, THAT INTELLIGENCE AND APTITUDE TESTS ADEQUATELY MEASURED CAPABILITIES IN PEOPLE TO PERFORM CERTAIN TASKS OR JOBS. HIS THESIS WAS THAT SUCH TESTS MEASURED, IF ANYTHING, CAPABILITIES TO PERFORM IN AN ACADEMIC SETTING AT ACADEMIC SKILLS. THEY DID NOTHING, HE CONTENDED, TO MEASURE THE ABILITY OF A POLICE CANDIDATE TO PERFORM POLICEMEN'S TASKS, FOR EXAMPLE. HE THEN WENT ON TO DESCRIBE THE CONCEPT THAT ONLY RANDOMLY SELECTED SKILLS REQUIRED OF PERSONS WHEN PERFORMING THE DESIRED JOB

COULD BE USED AS THE BASIS FOR A TEST OF APTITUDE FOR THAT JOB. HE CALLED THIS METHOD "CRITERION SAMPLING", AND IT FORMS THE DATA UPON WHICH VALID PREDICTORS OF FUTURE JOB COMPETENCY CAN BE ESTABLISHED. WHEN TESTS FOR THE PREDICTION OF FUTURE SUCCESS ON THE JOB HAVE BEEN DEVELOPED FROM SUCH DATA, THE VALIDITY COEFFICIENTS HAVE RISEN FROM AN AVERAGE OF 0.33 FOUND IN THE LITERATURE ON LEADERSHIP AND MANAGEMENT SKILL CHARACTERISTICS TO A MEAN OF 0.60. INASMUCH AS THE SQUARE OF A CORRELATION COEFFICIENT YIELDS THE PREDICTIVE VALIDITY OF A MEASURE, THE NEW JOB COMPETENCY ASSESSMENT RESULTS IN A THREE-FOLD IMPROVEMENT IN THE PREDICTIVE QUALITY OF THE MEASURE OVER PREVIOUS, TRADITIONAL MEANS. THE NAVY HAS THEREFORE CHOSEN TO DEVELOP SUCH A SET OF MEASURES TO DETERMINE THE LEADERSHIP AND MANAGEMENT QUALITIES OF ITS NON-COMMISSIONED AND COMMISSIONED OFFICERS. THIS IS PROBABLY AN APPROPRIATE TIME TO STATE A DEFINITION OF "COMPETENCY" AS USED IN THIS DISCUSSION, FOR IT WILL APPEAR MANY TIMES. "COMPETENCY" IS USED IN GOAL OR OUTCOME ORIENTED TRAINING TO IMPLY THE KNOWLEDGE, SKILLS, ABILITY, MOTIVES OR OTHER CHARACTERISTICS THAT CAN BE DEMONSTRATED TO RELATE DIRECTLY TO COMPETENT OCCUPATIONAL PERFORMANCE. IN THE PROGRAM NOW UNDER DISCUSSION THE ASSESSMENT OF THESE JOB RELATED COMPETENCIES

AMONG JOB INCUMBENTS IS BOTH THE SOURCE OF THE DETERMINATION OF THE COMPETENCIES RELATED TO SUPERIOR LEADERSHIP AND MANAGEMENT AND THE BASIS FOR THE MEASURES WHICH SUBSEQUENTLY ARE UTILIZED TO DETERMINE EACH STUDENT'S PRE- AND POST-TRAINING STATE.

THE JOB COMPETENCY ASSESSMENT PROCEDURE MAY BE DESCRIBED IN THREE STEPS, AS FOLLOWS:

1. THE IDENTIFICATION OF "SUPERIOR" AND "AVERAGE" CRITERION SAMPLES OF NAVY LEADERS. IN ORDER TO ACCOMPLISH THIS STEP, A SELECTED GROUP OF COMMANDING OFFICERS OF FLEET UNITS WERE ASKED TO IDENTIFY COMMISSIONED AND NON-COMMISSIONED OFFICERS ABOARD THEIR SHIPS WHO COULD BE PLACED IN EITHER A "SUPERIOR PERFORMER" CATEGORY OR AN "AVERAGE PERFORMER" CATEGORY.

2. THE CONDUCT OF "BEHAVIORAL EVENT" INTERVIEWS. THE OFFICERS AND PETTY OFFICERS IN THE SAMPLES TAKEN IN THE FIRST STEP WERE ASKED TO DESCRIBE IN BEHAVIORALLY SPECIFIC TERMS CRITICAL LEADERSHIP INCIDENTS IN WHICH THEY HAD BEEN INVOLVED. THE TECHNIQUE OF BEHAVIORAL EVENT INTERVIEWING, DEVELOPED BY DR. MCCLELLAND OF THE MCBER COMPANY, INVOLVES OBTAINING A NUMBER OF DESCRIPTIONS OF WHAT HE CALLS "BEHAVIORAL EPISODES." FOR EXAMPLE, A PERSON MIGHT BE ASKED TO DESCRIBE AN INCIDENT IN WHICH HE FELT

PARTICULARLY SUCCESSFUL (OR UNSUCCESSFUL) AND THEN TO DESCRIBE IN DETAIL THE EVENTS LEADING UP TO THE INCIDENT, WHEN AND WHERE IT OCCURRED, AND HOW HE WAS FEELING AND REACTING BEFORE, DURING AND AFTER THE INCIDENT. THE DISTINGUISHING ASPECT OF THIS INTERVIEW TECHNIQUE IS THAT IT ELICITS INFORMATION FROM WHICH ACTUAL BEHAVIORS CAN BE RECONSTRUCTED INSTEAD OF ELICITING INTERPRETATIONS OR RECOLLECTIONS OF GENERAL OUTCOMES.

3. THEMATIC CONTENT ANALYSIS OF BEHAVIORAL EVENTS. INASMUCH AS THE OFFICERS AND PETTY OFFICERS INTERVIEWED WERE SELECTED ACCORDING TO RATINGS BY THEIR COMMANDING OFFICERS, AND OTHER CRITERIA, IT IS POSSIBLE TO COMPARE THE "SUPERIOR" AND "AVERAGE" INTERVIEWERS IN TERMS OF THE CONTENT OF THEIR BEHAVIORAL EVENTS. THIS COMPARISON PROCESS INVOLVES, FIRST, THE IDENTIFICATION OF CHARACTERISTICS OR THEMES WHICH CAN BE DRAWN FROM THE RELATED INCIDENTS IN BOTH GROUPS OF THE SAMPLE. THEN IT INVOLVES THE DESIGN OF A SCORING SYSTEM WHICH WILL RELIABLY CREDIT A SET OF BEHAVIORAL EVENTS FROM AN INTERVIEW FOR THE PRESENCE OF THE THEMES OR CHARACTERISTICS. THESE THEMES WHICH ARE PRESENT IN THE EVENTS RELATED BY THE "SUPERIOR" GROUP AND NOT PRESENT IN THOSE RELATED BY THE "AVERAGE" GROUP BECOME THE COMPETENCY CHARACTERISTICS THAT ARE LIKELY TO LEAD TO HIGH PERFORMANCE. ONCE IDENTIFIED, THESE CHARACTERISTICS WERE SUBSEQUENTLY VALIDATED THROUGH TWO DIFFERENT MEANS.

THE THEMATIC CONTENT ANALYSIS FROM APPROXIMATELY EIGHT HUNDRED CRITICAL LEADERSHIP INCIDENTS COLLECTED FROM FLEET INTERVIEWS PROVIDED TWENTY-SEVEN RELIABLY DISTINGUISHABLE LEADERSHIP COMPETENCY CHARACTERISTICS, AND SUBSEQUENTLY DISTILLED DOWN TO FIVE MAJOR LEADERSHIP AND MANAGEMENT FUNCTIONS, OR FACTORS. STATISTICAL SUPPORT FOR THE DERIVATION OF THESE FIVE FACTORS MAY BE FOUND IN THE McBER AND COMPANY FINAL REPORT OF THE "ANALYSIS OF LEADERSHIP AND MANAGEMENT COMPETENCIES OF COMMISSIONED AND NON-COMMISSIONED NAVAL OFFICERS IN THE ATLANTIC AND PACIFIC FLEETS." HOWEVER, BECAUSE THESE FACTORS FORM THE BASIS FOR BOTH THE DETERMINATION OF THE BEHAVIORS TO BE TAUGHT IN THE TRAINING PROGRAM AS WELL AS THE MEASURES FOR DETERMINING THE EXTENT TO WHICH THE STUDENTS ACQUIRE SKILLS AND COMPETENCIES THROUGH THE TRAINING, IT IS IMPORTANT TO THIS DISCUSSION THAT THEY BE DESCRIBED, AND TO SOME EXTENT DISCUSSED.

THE TWENTY-SEVEN COMPETENCY CATEGORIES, AS DIVIDED INTO THE FIVE DISTINCT CONCEPTUAL CLUSTERS OR FACTORS ARE AS FOLLOWS:

FACTOR I: TASK ACHIEVEMENT

NAVY OFFICERS REPORTED NUMEROUS INCIDENTS IN WHICH THEY EXPRESSED CONCERN FOR ACHIEVEMENT, SET SPECIFIC GOALS, TOOK INITIATIVE TO SOLVE TECHNICAL PROBLEMS, OR COACHED OTHERS TO IMPROVE THEIR PERFORMANCE. FIVE COMPETENCY CATEGORIES MAKE UP THIS FACTOR:

1. CONCERN FOR ACHIEVEMENT: OFFICERS EXPRESSED A DESIRE TO "DO JOBS RIGHT", TO MEET STANDARDS OF EXCELLENCE, AND TO ADVANCE IN THEIR CAREERS, AND FELT PROUD WHEN THEY HAD DONE A JOB WELL.

2. TAKES INITIATIVE: OFFICERS DESCRIBED TAKING PERSONAL INITIATIVE TO OVERCOME OBSTACLES IN ACCOMPLISHING TASKS.

3. SETS GOALS: OFFICERS ARTICULATED SPECIFIC (OFTEN MEASURABLE), CHALLENGING BUT REALISTIC AND TIME-PHASED GOALS FOR THEIR OWN PERFORMANCE AND THAT OF THEIR SUBORDINATES AND UNIT.

4. COACHES: OFFICERS DESCRIBED HELPING SUBORDINATES TO ACCOMPLISH TASKS MORE EFFECTIVELY BY PROVIDING INFORMATION, "SHOWING THEM HOW", OR BY ENCOURAGING THEIR PERSONAL DEVELOPMENT THROUGH TRAINING OR OTHER ENRICHING EXPERIENCES.

5. <u>TECHNICAL PROBLEM SOLVING</u>: OFFICERS, PARTICULARLY ENLISTED PERSONNEL IN TECHNICAL RATES, DESCRIBED THINKING ANALYTICALLY IN SOLVING TECHNICAL PROBLEMS: OBSERVING DISCREPANCIES IN EQUIPMENT PERFORMANCE (PROBLEM FINDING), REASONING DEDUCTIVELY TO IDENTIFY THE CAUSES OF MALFUNCTIONS, LOCATING NEEDED RESOURCES, ANTICIPATING OBSTACLES, AND ACTING TO CORRECT PROBLEMS.

<u>FACTOR II: SKILLFUL USE OF INFLUENCE</u>

OFFICERS DESCRIBED BEING CONCERNED WITH INFLUENCE STRATEGIES -- PERSUASION, EXPLANATION, INSPIRATION, REWARDS -- TO ACCOMPLISH OBJECTIVES AND MOTIVATE SUBORDINATES TO WORK AS A TEAM. INFLUENCE SKILL IS AIDED BY CONCEPTUAL THINKING ABOUT SHORT- AND LONG-RANGE IMPACT AND BY EMOTIONAL SELF-CONTROL. FIVE COMPETENCY CATEGORIES WERE INCLUDED IN THIS FACTOR:

1. <u>CONCERN FOR INFLUENCE</u>: OFFICERS REPORTED BEING CONCERNED ABOUT INFLUENCING OTHERS ("I WANTED TO CONVINCE HIM"), USING THEIR OWN POWER IN INTERPERSONAL RELATIONS, AND BEING SENSITIVE TO THE POLITICAL FACTORS IN COMPLEX SITUATIONS.

2. <u>INFLUENCES</u>: OFFICERS DESCRIBED ACTING TO INFLUENCE OTHERS WITHOUT HAVING TO RESORT TO DIRECT ORDERS OR THREATS, USING INFLUENCE EFFECTIVELY TO ACHIEVE THEIR

ENDS WHILE MAKING OTHERS FEEL MORE EFFICACIOUS IN THE PROCESS.

3. CONCEPTUALIZES: OFFICERS DESCRIBED A HIGH LEVEL OF CONCEPTUAL ABILITY IN PROBLEM IDENTIFICATION, SYSTEMS ANALYSIS, AND POLICY FORMULATION. THIS COMPETENCY IS THE ABILITY TO SEE PATTERNS IN COMPLEX DATA, SEPARATE IMPORTANT INFORMATION FOR UNIMPORTANT, DEVELOP INTEGRATIVE CONCEPTS AND PRINCIPLES AND SUPPORT THESE WITH SPECIFIC DATA, AND RECONCILE EXCEPTIONS AND DISCREPANCIES, USUALLY WITH REGARD TO HAVING AN IMPACT ON OTHERS OR ON THE SYSTEM.

4. TEAM BUILDS: OFFICERS DESCRIBED ENCOURAGING SUBORDINATES TO WORK TOGETHER AS A TEAM, TO "BUY INTO" SHARED UNIT OR COMMAND PERFORMANCE GOALS, AND TO CREATE SYMBOLS AND EVENTS WHICH STIMULATED UNIT PRIDE AND IDENTITY.

5. REWARDS: OFFICERS REPORTED REWARDING OTHERS FOR GOOD TASK PERFORMANCE TO INFLUENCE AND MOTIVATE SUBORDINATES.

6. SELF-CONTROL: OFFICERS REPORTED CONSCIOUSLY CONTROLLING THEIR EMOTIONS, PARTICULARLY ANGER AND AFFILIATIVE TENDENCIES, WHEN THESE EMOTIONS THREATENED TO INTERFERE WITH THEIR ABILITY TO INFLUENCE OTHERS OR TO MAINTAIN AN EFFECTIVE LEADERSHIP ROLE.

FACTOR III: MANAGEMENT CONTROL

OFFICERS DESCRIBED USING A STRAIGHTFORWARD MANAGEMENT SEQUENCE OF PLANNING AND ORGANIZING, ISSUING DIRECTIONS, DELEGATING, MATCHING PEOPLE TO JOBS TO BE DONE, MONITORING RESULTS AND GIVING FEEDBACK IN MANY INCIDENTS. FIVE COMPETENCY CATEGORIES WERE INCLUDED IN THIS FACTOR:

1. PLANS AND ORGANIZES: OFFICERS REPORTED IDENTIFYING THE ACTIONS THEY NEEDED TO TAKE AT ONE POINT IN TIME TO ACHIEVE RESULTS AT SOME LATER TIME, SPECIFYING PERSONNEL, MATERIALS AND OTHER RESOURCES NEEDED, AND PRIORITIZING TASKS TO BE ACCOMPLISHED.

2. DIRECTS: OFFICERS, WHEN THEY DID NOT INFLUENCE SUBORDINATES, CLEARLY DIRECTED THEM TO PERFORM TASKS WITHOUT EXPLANATION AND IN THE ABSENCE OF PERSONALIZED THREATS OR PUNISHMENT.

3. DELEGATES: OFFICERS DESCRIBED CONSCIOUS USE OF THE CHAIN OF COMMAND TO GET SUBORDINATES TO TAKE RESPONSIBILITY FOR TASKS.

4. OPTIMIZES (PEOPLE-TASK): OFFICERS REPORTED REALISTICALLY ASSESSING PEOPLE IN MAKING PERSONNEL DECISIONS TO ASSIGN TASKS TO THOSE INDIVIDUALS MOST LIKELY TO DO THEM WELL, AND IN MAKING TRADE-OFFS BETWEEN TASK REQUIREMENTS AND INDIVIDUAL NEEDS (IN NAVY PARLANCE, "INTEGRATION OF MEN

AND MISSION").

5. <u>MONITORS RESULTS</u>: OFFICERS REPORTED MONITORING FOLLOWUP, CHECKING BACK TO SEE IF MANAGEMENT ACTIONS, SUBORDINATES OR EQUIPMENT IN FACT ACCOMPLISHED WHAT THEY WERE EXPECTED TO ACCOMPLISH IN A GIVEN TIME PERIOD.

6. <u>RESOLVES CONFLICTS</u>: OFFICERS DESCRIBED NEGOTIATING OR MEDIATING INTERPERSONAL DISPUTES TO A SUCCESSFUL RESOLUTION, DEFINED AS A "WIN-WIN" SOLUTION, IN WHICH BOTH PARTIES IN THE DISPUTE WERE RELATIVELY SATISFIED AND NEITHER LOST A DISPROPORTIONATE AMOUNT OF POWER, STATUS OR RESOURSES.

7. <u>GIVES FEEDBACK</u>: OFFICERS REPORTED GIVING SPECIFIC FEEDBACK TO SUBORDINATES ON THEIR TASK PERFORMANCE.

FACTOR IV: ADVISING AND COUNSELING

MANY LEADERS AND MANAGERS DESCRIBED LISTENING TO AND COUNSELING SUBORDINATES IN A HIGH PERCENTAGE OF THEIR CRITICAL LEADERSHIP INCIDENTS. COUNSELING INCIDENTS DEALT WITH FOUR ISSUES: PERFORMANCE, DISCIPLINARY MATTERS, PERSONAL PROBLEMS (INCLUDING DRUG, ALCOHOL, FINANCIAL AND FAMILY DIFFICULTIES) AND CAREER PLANNING. FOUR COMPETENCY CATEGORIES ARE INCLUDED IN THIS FACTOR:

1. LISTENS: OFFICERS REPORTED NOTICING WHEN SUBORDINATES APPEARED TO BE HAVING PROBLEMS, APPROACHING PEOPLE TO INVITE THEM TO TALK ABOUT ISSUES CONCERNING THEM, OR BEING PERCEIVED AS APPROACHABLE ("...THE BLACK GUYS CAME TO SEE ME BECAUSE, THEY SAID, THEY FELT I WAS THE ONLY ONE THEY COULD TALK TO.").

2. UNDERSTANDS: OFFICERS DESCRIBED BEING ABLE TO "HEAR WHAT OTHERS ARE TRYING TO SAY" (ACCURATE EMPATHY OR INSIGHT INTO SUBORDINATES' NEEDS, MOTIVES OR HIDDEN AGENDA).

3. HELPS: OFFICERS DETAILED THE ACTIONS THEY TOOK TO HELP SUBORDINATES IN COUNSELING SITUATIONS, INCLUDING GIVING ADVICE, MAKING TIME AVAILABLE TO TALK, ACTING DIRECTLY TO "FIGHT FOR THEIR PEOPLE," OR MAKING APPROPRIATE REFERRALS TO SOURCES OF HELP (MEDICAL PERSONNEL, CHAPLAINS, DRUG AND ALCOHOL TREATMENT FACILITIES).

4. POSITIVE EXPECTATIONS: OFFICERS EXPRESSED POSITIVE EXPECTATIONS OF AND REGARD FOR THEIR SUBORDINATES.

FACTOR V: COERCION

NAVY OFFICERS DESCRIBED CRITICAL INCIDENTS IN WHICH THEY USED RANK OR THREATS TO MOTIVATE OTHERS TO ACT, EXPRESSED NEGATIVE EXPECTATIONS OF SUBORDINATES, DISCIPLINED THEM, ACTED IMPULSIVELY, AND RESOLVED CONFLICTS BY FORCE

OR FAILED TO RESOLVE THEM. FIVE COMPETENCY CATEGORIES WERE INCLUDED IN THIS FACTOR:

1. COERCES: OFFICERS DESCRIBED USING RANK AND BOTH GENERAL AND PERSONALIZED THREATS TO MOTIVATE SUBORDINATES.

2. NEGATIVE EXPECTATIONS: OFFICERS EXPRESSED NEGATIVE REGARD AND EXPECTATIONS FOR THEIR SUBORDINATES ("HE IS NO GOOD -- THERE'S NO WAY HE'S GOING TO MAKE IT").

3. DISCIPLINES: OFFICERS DESCRIBED PUNISHING SUBORDINATES BY GIVING THEM NEGATIVE FEEDBACK, POOR FITNESS OR EVALUATION REPORTS, OR USING STANDARD UCMJ PROCEDURES.

4. ACTS IMPULSIVELY: OFFICERS REPORTED EXPRESSING THEIR EMOTIONS WITHOUT INHIBITION -- PRIMARILY ANGER ("I BLEW MY STACK AT HIM") AND OCCASIONALLY AFFILLIATION ("MY BUDDIES HAD TO COME FIRST").

5. FAILS TO RESOLVE CONFLICTS: OFFICERS DESCRIBED SITUATIONS IN WHICH THEY DID NOT REACH "WIN-WIN" RESOLUTIONS OF CONFLICTS, EITHER BECAUSE THEY RESOLVED CONFLICTS BY UNILATERAL FORCE OR BY AVOIDING DEALING WITH THE CONFLICT.

CONSTRUCT VALIDATION FOR THESE FACTORS AND SUPPORTING CATEGORIES IS AVAILABLE IN THE LITERATURE, AND FOR THOSE WHO WISH TO PROBE DEEPER INTO THIS ASPECT OF THE COMPETENCIES, THE MCBER REPORT MENTIONED EARLIER CITES THE RELEVANT REFERENCES.

AFTER THESE COMPETENCY FACTORS AND CATEGORIES WERE DERIVED FROM THE ANALYSIS OF THE DATA GATHERED FROM THE INTERVIEWS IN BOTH THE PACIFIC AND ATLANTIC FLEETS, THEY WERE CROSS VALIDATED IN FOLLOW-ON INTERVIEWS IN BOTH FLEETS TO DETERMINE THE EXTENT TO WHICH THE CRITERIA COULD DISCRIMINATE BETWEEN "SUPERIOR" AND "AVERAGE" COMMISSIONED AND NON-COMMISSIONED OFFICERS DRAWN FROM EACH FLEET. THE FIRST FOUR FACTORS, THAT IS, TASK ACHIEVEMENT, SKILLFUL USE OF INFLUENCE, MANAGEMENT CONTROL AND ADVISING AND COUNSELING, CLEARLY DISTINGUISHED BETWEEN THE TWO CATEGORIES OF OFFICERS AND PETTY OFFICERS, ALTHOUGH THE FACTOR CALLED COERCION FAILED TO SO DISCRIMINATE, IT APPEARING EQUALLY AS A CHARACTERISTIC OF BOTH "SUPERIOR" AND "AVERAGE" PERSONNEL.

PERHAPS THIS IS AN APPROPRIATE PLACE TO ADDRESS THE QUESTION OF THE DIFFERENCE, IF ANY, BETWEEN JOB COMPETENCY ASSESSMENT AND JOB/TASK ANALYSIS. IT WOULD SEEM THAT THEY BOTH ACCOMPLISH A COMMON RESULT...THEY PROVIDE A DATA BASE FROM WHICH INSTRUCTIONAL PROGRAMS AND ACHIEVEMENT MEASURES

MAY BE DERIVED, AND THEY PROVIDE A MEANS BY WHICH WE MAY PRIORITIZE SKILLS AND KNOWLEDGES REQUIRED FOR THE SUCCESSFUL ACCOMPLISHMENT OF A JOB SO THAT WE MAY PLACE OUR TRAINING RESOURCES WHERE THEY WILL DO THE MOST GOOD. BUT THERE ARE ALSO SOME VERY IMPORTANT DIFFERENCES BETWEEN THEM, IT SEEMS TO US, AND THESE SHOULD BE UNDERSTOOD. FIRSTLY, AND MOST IMPORTANTLY, JOB/TASK ANALYSIS PROVIDES INFORMATION AS TO WHAT MUST BE ACCOMPLISHED IN ORDER TO PERFORM A SPECIFIC TASK OR JOB, WHEREAS JOB COMPETENCY ASSESSMENT PROVIDES INFORMATION ABOUT HOW TO GET THE JOB DONE IN THE BEST AND MOST EFFECTIVE FASHION. SECONDLY, WHILE BOTH ADDRESS THE COGNITIVE DOMAIN, JOB/TASK ANALYSIS DEALS PRIMARILY WITH THE PSYCHOMOTOR DOMAIN WHILE JOB COMPETENCY ASSESSMENT DEALS PRIMARILY WITH THE AFFECTIVE DOMAIN. AND THIRDLY, JOB/TASK ANALYSIS MOST OFTEN CONCERNS ITSELF WITH TECHNICAL MATTERS, WHILE JOB COMPETENCY ASSESSMENT CONCERNS ITSELF WITH NON-TECHNICAL MATTERS.

AS WE HAVE DISCUSSED BRIEFLY, THE JOB COMPETENCY ASSESSMENT FUNCTION HAS PROVIDED US WITH A DATA BASE UPON WHICH TO CONSTRUCT THE CURRICULUM FOR THE TRAINING PROGRAM IN LEADERSHIP AND MANAGEMENT SKILLS. IT HAS ALSO PROVIDED US AT THE SAME TIME WITH THE DATA BASE FROM WHICH TO DEVELOP INSTRUMENTS FOR THE MEASUREMENT OF PRE- AND POST-TRAINING SKILLS IN LEADERSHIP AND MANAGEMENT.

THE DEVELOPMENT AND PACKAGING OF A SET OF ASSESSMENT INSTRUMENTS TO MEASURE THE COMPETENCIES IDENTIFIED AS THOSE DISTINGUISHING BETWEEN SUPERIOR AND AVERAGE LEADERS AND MANAGERS IS CALLED THE NAVY LEADERSHIP AND MANAGEMENT SKILLS TEST BATTERY. IT HAS BEEN ADMINISTERED TO APPROXIMATELY ONE THOUSAND COMMISSIONED AND NON-COMMISSIONED OFFICERS AT EIGHT LEADERSHIP LEVELS, ALTHOUGH AT THE TIME OF THIS PRESENTATION THE DATA ARE NOT YET AVAILABLE FOR REPORTING. THE LEVELS AT WHICH THE BATTERY WAS ADMINISTERED WERE PETTY OFFICER, LEADING PETTY OFFICER, CHIEF PETTY OFFICER, MASTER CHIEF PETTY OFFICER, DIVISION OFFICER, DEPARTMENT HEAD, EXECUTIVE OFFICER AND COMMANDING OFFICER.

THE PURPOSES OF THIS TEST BATTERY DEVELOPMENT WERE AS FOLLOWS:

1. TO MEASURE THE COMPETENCIES IDENTIFIED AS DISTINGUISHING BETWEEN SUPERIOR AND AVERAGE NAVY LEADERS AND MANAGERS.

2. TO DEVELOP TESTS WHICH MEET THE AMERICAN PSYCHOLOGICAL ASSOCIATION PSYCHOMETRIC STANDARDS FOR RELIABILITY AND VALIDITY.

3. TO DEVELOP TESTS THAT ARE NAVY RELEVANT AND FACE VALID.

4. TO DEVELOP TESTS THAT ARE EASILY ADMINISTERED AND SCORED.

NOW INASMUCH AS COMPETENCIES DESCRIBE ATTITUDES PEOPLE HAVE OR ACTIONS THEY TAKE, INSTEAD OF WHAT THEY KNOW, TESTS MUST MEASURE WHAT PEOPLE DO, NOT WHAT THEY KNOW. UNFORTUNATELY, MOST TESTS AVAILABLE HAVE BEEN DESIGNED TO MEASURE KNOWLEDGE. SO A NEW APPROACH TO THE DESIGN OF TESTS HAD TO BE MADE, AND THIS WAS DONE UNDER THE GUIDANCE OF DR. MCCLELLAND OF THE MCBER COMPANY. THESE TESTS HAVE THE FOLLOWING ATTRIBUTES:

1. THEY ASSESS COMPETENCIES, THAT IS, THOSE BEHAVIORS THAT DISTINGUISH BETWEEN AVERAGE AND SUCCESSFUL PERFORMANCE INVOLVED IN CLUSTERS OF LIFE SKILLS,

2. THEY SHOULD BE DEVELOPED BY EXAMINING THE BEHAVIORS OF PEOPLE EXHIBITING THOSE BEHAVIORS TO BE MEASURED,

3. THEY SHOULD TAP OPERANT AS WELL AS RESPONDENT THOUGHT. THE RESULTING BATTERY ARE A COMPOSITE OF FOUR TESTS WHICH WERE AVAILABLE IN THE MARKET, AND FOUR WHICH REQUIRED DEVELOPMENT BY THE MCBER TEST DESIGNERS. TABLE I IS A DESCRIPTION OF THESE TESTS BY FUNCTION AND NAME, AND INDICATES WHICH ARE THOSE AVAILABLE FROM STANDARD TESTS AVAILABLE IN THE MARKET, AND WHICH ARE THE RESULT OF MCBER

DEVELOPMENT. THE OUTCOME OF THE PILOT TESTS OF THESE MEASURES, AND A MORE ELABORATE AND COMPLETE DISCUSSION OF THE CONSTRUCTION OF THE NEW MEASURES MAY BE OBTAINED FROM THE CHIEF OF NAVAL PERSONNEL IF DESIRED.

TABLE I

TESTS, THEIR FORMAT, AND COMPETENCIES TESTED FOR IN THE NAVY LEADERSHIP AND MANAGEMENT SKILLS TEST BATTERY

TESTS WHICH EXIST

TEST	FORMAT	COMPETENCIES
STRONG-CAMPBELL INTEREST INVENTORY	RATE YOUR PREFERENCES ON A NUMBER OF DIMENSIONS	TECHNICAL PROBLEM SOLVING
PICTURE STORY EXERCISE	WRITE IMAGINATIVE STORIES TO EACH OF SIX PICTURES	CONCERN FOR ACHIEVEMENT TAKES INITIATIVE HAS CONCERN FOR INFLUENCE. HAS SELF-CONTROL ACTS IMPULSIVELY

TESTS WHICH EXIST (CONTINUED)

TEST	FORMAT	COMPETENCIES
WORK ANALYSIS QUESTIONNAIRE	EXPRESS YOUR WORK PREFERENCES ON A SERIES OF STATEMENTS	CONCERN FOR ACHIEVEMENT CONCERN FOR INFLUENCE
ORGANIZATIONAL CLIMATE SURVEY QUESTIONNAIRE	DESCRIBE AN IDEAL WORK CLIMATE FOR YOU BY RESPONDING TO A SERIES OF STATEMENTS	CONCERN FOR ACHIEVEMENT CONCERN FOR INFLUENCE

TESTS WHICH WERE DEVELOPED

TEST	FORMAT	COMPETENCIES
LISTENING AND COUNSELING, PART I	WRITE OUT ANSWERS ABOUT TAPED VOICE EXCERPTS AND ABOUT PICTURES	LISTENS UNDERSTANDS HELPS
LISTENING AND COUNSELING, PART II	ANSWER QUESTIONS AFTER HEARING EACH OF FOUR MONOLOGUES	LISTENS UNDERSTANDS HAS POSITIVE EXPECTATIONS

TESTS WHICH WERE DEVELOPED (CONTINUED)

TEST	FORMAT	COMPETENCIES
MANAGEMENT OF PROBLEMS TEST	CHOOSE THE THREE PREFERRED RESPONSES OUT OF SIX RESPONSES GIVEN FOR EACH OF TWENTY SITUATIONS.	INFLUENCES TEAM BUILDS DIRECTS (AUTHOI-TARIAN) RESOLVES CON-FLICTS COERCES FAILS TO RESOLVE CONFLICTS
MANAGERIAL STYLE QUESTIONNAIRE	GIVE EACH OF THREE RESPONSES TO A SER-IES OF SENTENCES A PREFERRED NUMBER OF POINTS	COACHES REWARDS PLANS AND ORGANIZES DELEGATES MONITORS RESULTS GIVES FEEDBACK PUNISHES
PLANNING EXERCISE	FIND THE MOST EFFICIENT SCHEDULE FOR A SERIES OF TASKS	PLANS AND ORGANIZES

To summarize, we may say that for the first time in the Navy there has been described through valid research a set of competencies, that is, skills and characteristics, which are demonstratably capable of discriminating between superior and average leaders and managers among naval officers and petty officers, and which can therefore be reliably used as a data base for the development of training programs in leadership and management, including the instruments for the measurement of achievement of the required behaviors for good leadership and management in the Navy. The Bureau of Naval Personnel, through the work of their contractor, the McBer and Company of Boston, are currently developing leadership and management training programs at a number of accession points and grade levels of naval personnel, and the Naval Education and Training Command, to be later charged with the responsibility for the implementation of these programs, is carefully monitoring the process. Initial implementation is planned to commence in 1978.

TESTING IN THE AFMET PROGRAM

Wallace Bloom, Ph.D.
Wilford Hall USAF Medical Center

Introduction: Premature termination of military service is often caused by psychological problems that existed prior to service and led to poor emotional adaptation to military life. There has been previous research towards the development of an objectively scored questionnaire capable of discriminating between individuals who will develop emotional or characterological difficulties during training and those who will successfully adjust to military life. Danielson and Clark (1954) studied a sample of 15,550 Army recruits; Jensen's (1964) 82-item questionnaire had been given to 9,194 Air Force recruits, and Plag (1962) used a 195-item questionnaire on 20,000 Navy recruits. LaChar, et al (1974) reported on a 1972 investigation of 14,804 male Air Force recruits and the evolution of the history-opinion-interest from (HOI). Guinn, et al (1975) concluded that the HOI has some practical usefulness as a rough preliminary screening device. They recommended that use of this screening device be limited to preliminary screening only and that addition psychometric and/or psychiatric assessment be mandatory before any personnel action is recommended.

A proposal for further research and development of a military adaptability screening test was prepared in 1974 by Captain Charles I. Bisbee, Major George E. Hargrave, and Colonel John C. Sparks. The research protocol designed by the Department of Mental Health, Wilford Hall USAF Medical Center, in conjunction with the Surgeon, Air Training Command and other ATC agencies, was implemented in June 1975 and identified as AFMET (Air Force Medical Evaluation Test Program). (Bloom, 1977a)

PHASE I: On the arrival day, each airman entering the Air Force was given a psychological test. Those identified by computer scoring as low mental health risks (approximately 93%) continued basic training without further evaluation.

PHASE II: By the first day of training, those not identified as low-risk were called in for individual mental status interviews with mental health technicians and additional psychological tests are given. The reports of these interviews and tests are reviewed by a senior clinical psychologist who determined which additional trainees will now be identified as low-risk and did not require further evaluation.

Presented at the 19th Conference of the Military Testing Association, October 18, 1977, San Antonio, Texas.

PHASE III: Usually by the fifth day of training, those airmen (approximately 2 to 3% of the trainees) not already identified as low-risk were referred to the Mental Hygiene Clinic for further evaluations. They brought with them reports of behavior observations and comments from their training squadrons (ATC Form 582). Clinical mental health interviews were conducted and additional psychological tests are selected and given on an individual basis as necessary. Those without serious problems were returned to duty, and the few identified as being psychotic were referred to the hospital for treatment and further action. Some showed evidence of specific character and behavior disorders of such a nature as to seriously impair military performance. They were referred back to their training squadron commanders with recommendations for administrative separation with medical classification and behavior reports. A few are recommended for referral to other special agencies. (See Illustration 1).

HOI: Phase I testing was limited to the original History Opinion Inventory (HOI) composed of 100 items that the subject was to identify as True or False as applied to him. Examples are: I quit school because I was failing. I was active in sports during high school. I have been in trouble with the police. I like hunting very much. Marking was on opscan type answer sheets which were scored by computer and the names and other identification of those selected for further testing were contained in the printout sheets.

In 1975, during the initial days of the project, it was quickly learned that the HOI Adaptation Index cutoff score of 12 that had identified 12% of the sample in 1972 only identified ½ of 1% in 1975. It was believed that due to the termination of the Draft Law and other factors, the new recruits were scoring significantly different that the earlier subjects. The cutoff score was reduced 4 points. 38,529 basic trainees took the HOI between June 1 and November 7, 1975, and approximately 6% were identified for Phase II screening. (Table 1).

The HOI has two subscales. PEI (Emotional Instability) consists of 18 scored items and the PDA (Drug Use Admission of 26 scored items. Nine of the items are critical on both subscales. In November 1975, the test was reduced from 100 items to 50 without changing or eliminating any of the scored items. The five-page booklet was replaced by a one-page card which could be aligned with the answer sheet to minimize marking errors. The cutoff score remained unchanged.

On 1 October 1976, the AFMET program shifted from a Research Project to a Standard Operating Procedure. The HOI test was given after the first day of training rather than on arrival. The trainees would have spent at least one night in their dorms, and met

PHASE III: Usually by the fifth day of training, those airmen (approximately 2 to 3% of the trainees) not already identified as low-risk were referred to the Mental Hygiene Clinic for further evaluations. They brought with them reports of behavior observations and comments from their training squadrons (ATC Form 582). Clinical mental health interviews were conducted and additional psychological tests are selected and given on an individual basis as necessary. Those without serious problems were returned to duty, and the few identified as being psychotic were referred to the hospital for treatment and further action. Some showed evidence of specific character and behavior disorders of such a nature as to seriously impair military performance. They were referred back to their training squadron commanders with recommendations for administrative separation with medical classification and behavior reports. A few are recommended for referral to other special agencies. (See Illustration 1).

HOI: Phase I testing was limited to the original History Opinion Inventory (HOI) composed of 100 items that the subject was to identify as True or False as applied to him. Examples are: I quit school because I was failing. I was active in sports during high school. I have been in trouble with the police. I like hunting very much. Marking was on opscan type answer sheets which were scored by computer and the names and other identification of those selected for further testing were contained in the printout sheets.

In 1975, during the initial days of the project, it was quickly learned that the HOI Adaptation Index cutoff score of 12 that had identified 12% of the sample in 1972 only identified ½ of 1% in 1975. It was believed that due to the termination of the Draft Law and other factors, the new recruits were scoring significantly different that the earlier subjects. The cutoff score was reduced 4 points. 38,529 basic trainees took the HOI between June 1 and November 7, 1975, and approximately 6% were identified for Phase II screening. (Table 1).

The HOI has two subscales. PEI (Emotional Instability) consists of 18 scored items and the PDA (Drug Use Admission of 26 scored items. Nine of the items are critical on both subscales. In November 1975, the test was reduced from 100 items to 50 without changing or eliminating any of the scored items. The five-page booklet was replaced by a one-page card which could be aligned with the answer sheet to minimize marking errors. The cutoff score remained unchanged.

On 1 October 1976, the AFMET program shifted from a Research Project to a Standard Operating Procedure. The HOI test was given after the first day of training rather than on arrival. The trainees would have spent at least one night in their dorms, and met

their primary instructors prior to taking this test. During a few trial days, it was found that the cutoff score of 8 points now identified 22% instead of the approximate 6 to 7% identified from June 1, 1975 to June 15, 1976. The cutoff score was raised to 11 points with daily monitoring to shift it to 12 when the quota might be exceeded.

MMPI: The selectees' MMPI responses were marked on Optiscan sheets, machine scored, and computer print-outs were returned a day later. These print-outs were based on a program developed by LaChar (1974). The consolidated data (Standardized Report of Interview (SRI) and MMPI) were then reviewed by a clinical psychologist who determined which "low risk" individuals warranted continuation in training without referral to the Mental Hygiene Clinic for further action. The others were referred to that clinic.

Approximately 400 subjects were given MMPI's each month (Bloom, 1977b). As the program continued, it was noted that a significant number of airmen scored highly elevated MMPI T-scores, often above 100, but were reported as within normal limits by the interviewers. During the first two months of the study, many airmen with these elevated MMPI scores, but low risk interview reports, were referred to the Mental Hygiene Clinic. Further psychiatric assessment there usually resulted in these airmen's continuation in basic training. The MMPI norms used were based on the general Minnesota normal sample (Dahlstrom, et al, 1960, pp 437-8). During July, the MMPI's of 17 year-old airmen were recorded by hand using the T-score conversions for basic scales without K corrections of Minnesota Adolescents, age 17 (Dahlstrom, et al, 1972, pp 397-99). Although extreme T-scores were somewhat less elevated than when converted the traditional way, it was questioned whether these norms were appropriate. It was hypothesized that neither the traditional adult nor adolescents, age 17 norms were relevant for judging the MMPI responses of basic airmen.

Subjects: Both male and female recruits in basic training squadrons were tested rather than the HOI selectees which might have been a skewed sample. Entire squadrons were tested during the period August to October, 1975. Personnel came from all over the United States. Data was obtained from 1152 males and 805 females.

Procedures: Standard MMPI booklets were used and responses were marked on Opscan sheets for subsequent machine scoring. Raw scores for each category were punched on IBM cards and the data statistically analyzed by the Biometrics Division of the Air Force Systems Command. The analysis of variance for disproportionate data using a general regression model was based on Graybill's work.

Results: The demographic composition of the Lackland population reflected recruit criteria that each must be high school graduates or pass a GED test equivalent. Only 3.4% of the men and 2.8% of the women were not high school graduates. 12.2% of the males and 18.9% of the females had reported education beyond high school level (See Table 2). Over eighty (80%) percent were "Anglo-Americans" (Caucasian), and about ten (10) percent of the total sample were "Black-Americans". Further details are shown on Table 3. Only two-percent of the men, but almost 13 percent of the women were over 23 years old (See Table 4).

Means and Standard Deviations on Basic MMPI factors were obtained for 1152 male and 805 female participants (See Table 5). Comparisons with traditional MMPI norms for adults revealed statistically significant differences in all scaled except Social Introversion (0-Si). A second set of analyses between the Lackland population and the norms of Minnesota Normal Adolescents, age 17, indicated significant differences in almost all scales (except Males: Lie, Depression, and Hysteria and except Females: Lie and Paranoia).

Conclusions: Norms for purposes of comparison must be relevant and will be meaningless or even misleading if they are not based on groups of people with whom it is sensible to compare the individuals we are psychologically assessing. Neither the MMPI norms for Minnesota Adults nor for Minnesota Normal Adolescents age 17, were relevant for comparison with Air Force Basic Trainees.

The standard MMPI norms were not designed for use with a young adult military population and characteristics cited by Hathaway and Briggs (1957) follow:

VARIABLES	LACKLAND	MINNESOTA
Geographic	National	Regional
Age	Predominately 17-23	Mean 33
Race	Included minorities	White
Education	12 or more	Mean 9.7 to 10.0
Population		
Male	1152	111 to 345
Female	805	118 to 397
Note		varied by scale
Tested	1975	1940

It was appropriate to establish new norms based on these individuals and this has been done along with appropriate T-conversions (Tables 6 and 7) and plotting charts (Illustrations 2,3).

From November 15, 1975 to April 1, 1976 these Lackland MMPI norms were utilized and along with a sentence completion test as part of Phase II testing. The MMPI was not used from April 1, 1976 to June 15, 1976 and fewer unnecessary referrals were made to Phase III, the Mental Hygiene Clinic. The AFMET '77 program does not use the MMPI for Phase II but it is used in some selected individual cases at the Mental Hygiene Clinic for Phase III. Formerly one out of six Phase III selectees were discharged and currently about one of three. The MMPI proved ineffective for Phase II testing.

BSCS: In November 1976, as part of Phase II interviewing the Bloom Sentence Completion Survey, (BSCS) was given to facilitate establishing rapport between the trainees and the enlisted mental health technicians conducting the interviews. Initially intended to serve as an icebreaker for the interview and to provide some advance information, it was found to take only about 12 minutes for group testing. The results were read by the interviewer before the trainee was seen and some key responses were underscored and often referred to during the interview. The interviewers later became interested in numerical scoring and were instructed in identifying each response as positive, neutral, or negative by categories.

This test (Bloom, 1975) purports to indicate both positive and negative aspects of attitudes towards: people, physical self, family, psychological self, self-directedness, work, and accomplishment in addition to identifying some irritants in each subject's life. After a week the inter-scorer correlations were about .90 with rarely more than 4 items scored differently than by the instructor. After a month, it was rare for them to differ on more than two of the 35 scored responses. Scoring time was usually less than seven minutes. For comparison purposes the scores of random trainees were compiled. These mean scores and standard deviations are in Table 8.

The 2879 trainees interviewed at Lackland AFB during Phase II of the AFMET research program (January 1, 1976 - June 15, 1976) had: medical record notes, a standardized report of their interviews on an opscan type sheet (SRI), a scored sentence completion test, and in most cases (all those from 2 January 1976 to 1 April 1976) the computer print-out of each MMPI. 2213 of these trainees were returned to duty in their training squadrons while 666 were scheduled for further evaluations at the Mental Hygiene Clinic. As noted on Table 9, the BSCS composite scores of the returned to duty trainees averaged 8.8450 while that of those referred to Mental Hygiene Clinic for further evaluation was -.0376. There were statistically significant differences on each subscore between the group cleared and those to be further evaluated. These differences as reflected in the T-scores were so large that the chances of being accidental were

less than one in a thousand. Of the 666 individuals referred to Phase III, 522 who were subsequently returned to duty (RTD) had an average composite score of .3008 while 114 who were discharged had an average composite score of minus 2.0364. The subscores for attitudes towards physical self, self-directedness and work were significantly different.

Conclusions: Of the psychological tests used in the program, the HOI functioned about as well as predicted as it identified approximately one-third of the basic trainees given mental health discharges. The other two-thirds were referred by the training squadrons directly to the Mental Hygiene Clinic usually after the 10th day of training. The recommendations made about the HOI in 1975 (Guinn et al) are still pertinent with regard to need for revalidation, special scales for a WAF population and its use in combination with additional aptitudinal and biographic data available on all recruits.

The MMPI did not work out in AFMET as a screening device at Phase II level but with current Air Force norms appears to have use along with other tests on a selected individual basis as part of diagnosis in Phase III.

The Bloom Sentence Completion Survey proved useful clinically as part of the Phase II interviews and assessment. Statistically it differentiated between groups of trainees who did or did not require further evaluations and between those discharged or returned to duty. Further research need be undertaken regarding the composite score of the seven subtests.

Additional research on the HOI is being conducted by the Personnel Division, Headquarters USAF which has given a version of the HOI to over 80,000 subjects at the AFIE stations and by analysis of earlier Lackland data by the Human Resources Lab. Further research to follow trainees through their enlistments, identify those with early separations for mental health associated causes and compare their AFMET test scores, still remains to be done and the first of the AFMET subjects will not complete their enlistments until 1979.

During the research year of the AFMET program (June 1, 1975 to June 15, 1976) 80,732 new arrivals were given the HOI through Phase I of the program; of these 5369 received Phase II interviews and tests which resulted in 1331 being referred to Phase III for further evaluations. Four hundred and forty-four (444) of these basic trainees were discharged (including 79 by squadrons, 59 for medical reasons, and 306 recommended by Mental Health Clinic).

During the operational period October 1, 1976 to October 1, 1977, reportedly 73,666 went through Phase I; 4918 were selected for Phase II; and, 1054 were referred to Phase III. Five hundred

and twelve (512) were recommended for discharge through the Mental Hygiene Clinic (Phase III) and 308 Phase II selectees were discharged by other agencies before the AFMET actions could be completed. Seventeen were discharged by the hospital for mental health reasons.

The early identification of trainees with significant psychological problems facilitated their return to their civilian lives without further trauma due to military stress and to termination of further Air Force investment in their training. It has been estimated that well over one million dollars has been saved at Lackland each year of the AFMET program plus further indirect savings by curtailment of investments in technical school training for individuals who do not complete normal enlistments. The AFMET program seems to have benefited both the USAF and the individuals.

REFERENCES

Bloom, W., Manual, Bloom Sentence Completion Test; Student and Adult. Effectiveness Training Associates of San Antonio, 133 Twinleaf Ln., San Antonio, Texas 78213, 1975.

Bloom, W., Air Force Medical Evaluation Test. Medical Service Digest, United States Air Force 28(2) Mar/Apr 1977, 17-20.

Bloom, W., Relevant MMPI Norms for Young Adult Air Force Trainees. Scheduled for publication in Journal of Personality Assessment in December, 1977.

Danielson, J.R. & Clark, J.H. A personality inventory for induction screening. Journal of Clinical Psychology, 1954, 10, 137-143.

Guinn, N., Johnson, A.L. & Kantor, J.E. Screening for adaptability to military service. Report No. AFHRL-TR-75-30, Air Force Resources Laboratory, May 1975.

Jensen, M.B., Adjustive and non-adjustive reactions to basic training in the Air Force. Journal of Social Psychology, 1961, 55, 33-41.

LaChar, D., The MMPI, Clinical Assessment and Automated Interpretation, Los Angeles, California, Western Psychological Services, 1974.

LaChar, D., Sparks, J.C. & Larsen, R.M., Psychometric prediction of adaptation for USAF basic trainees. Journal of Community Psychology, 1974, 2(3), 268-277.

Plag, J.A. Pre-enlistment variables related to the performance and adjustment of Navy recruits. Journal of Clinical Psychology, 1962, 19, 168-171.

LIST OF ILLUSTRATIONS & TABLES

ILLUSTRATION I

AFMET-1977

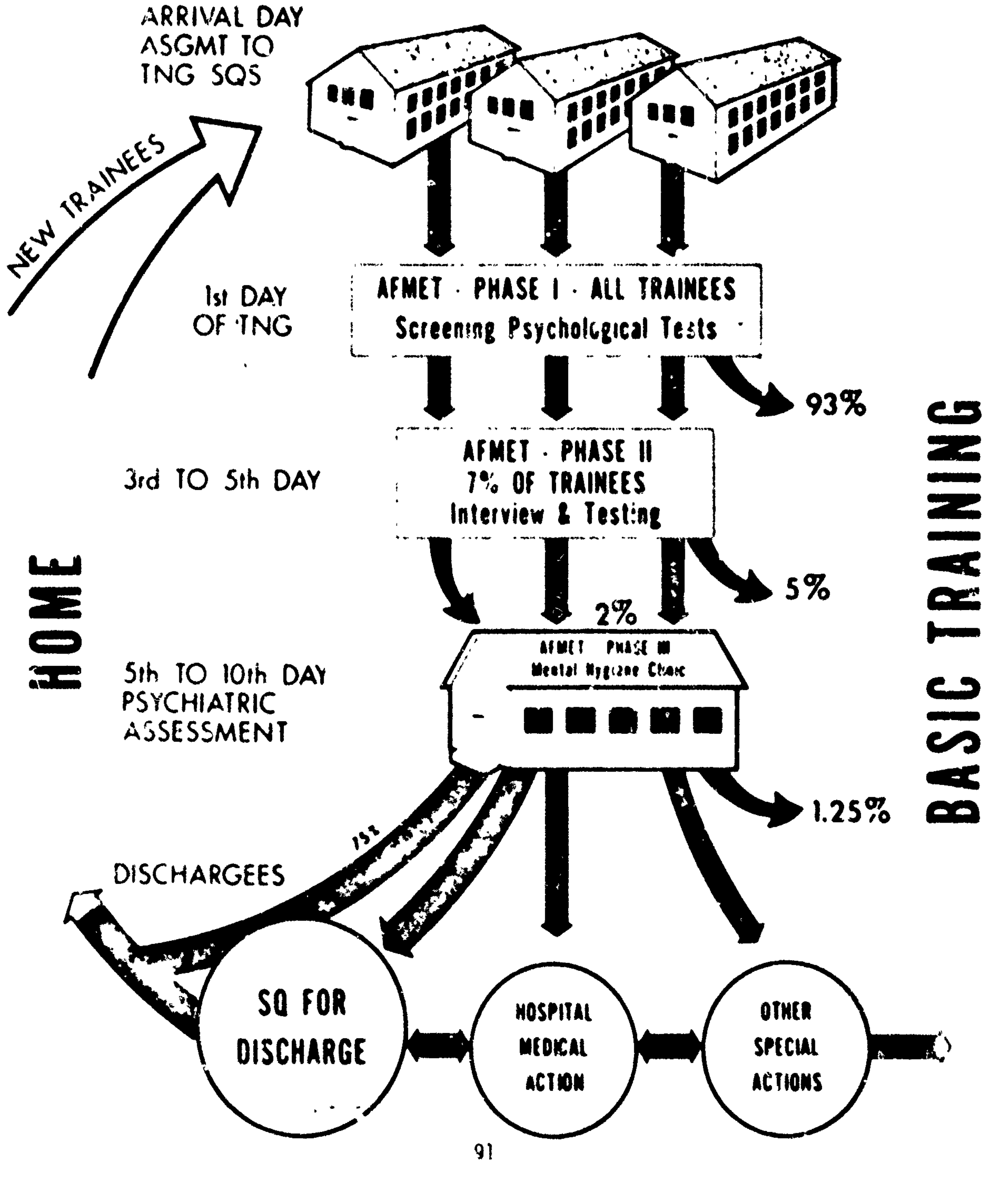
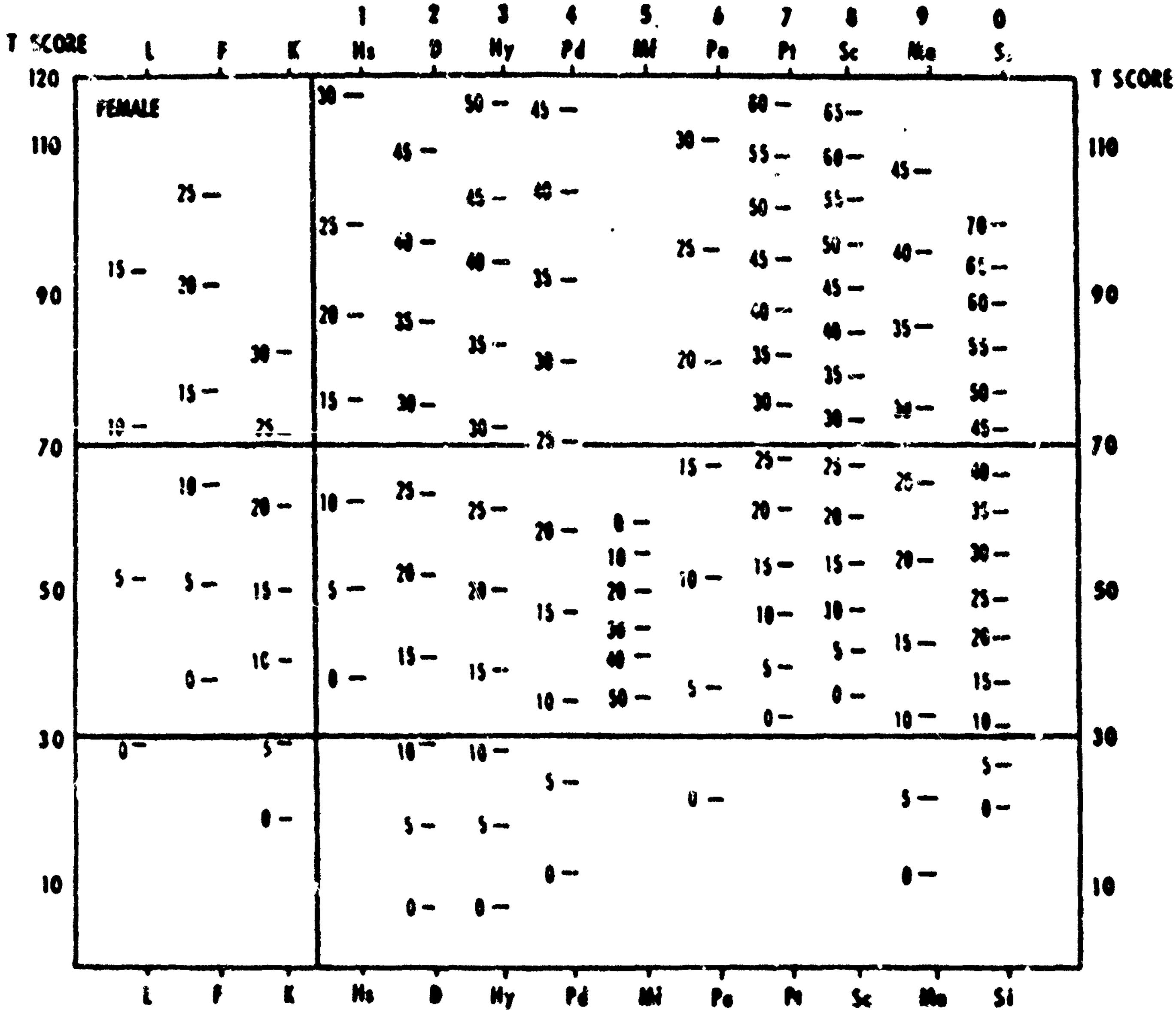


ILLUSTRATION 2

FEMALE

T SCORE CONVERSIONS FOR MMPI SCALES WITHOUT K CORRECTIONS.
BASED ON 805 FEMALE USAF BASIC TRAINEES, 1975



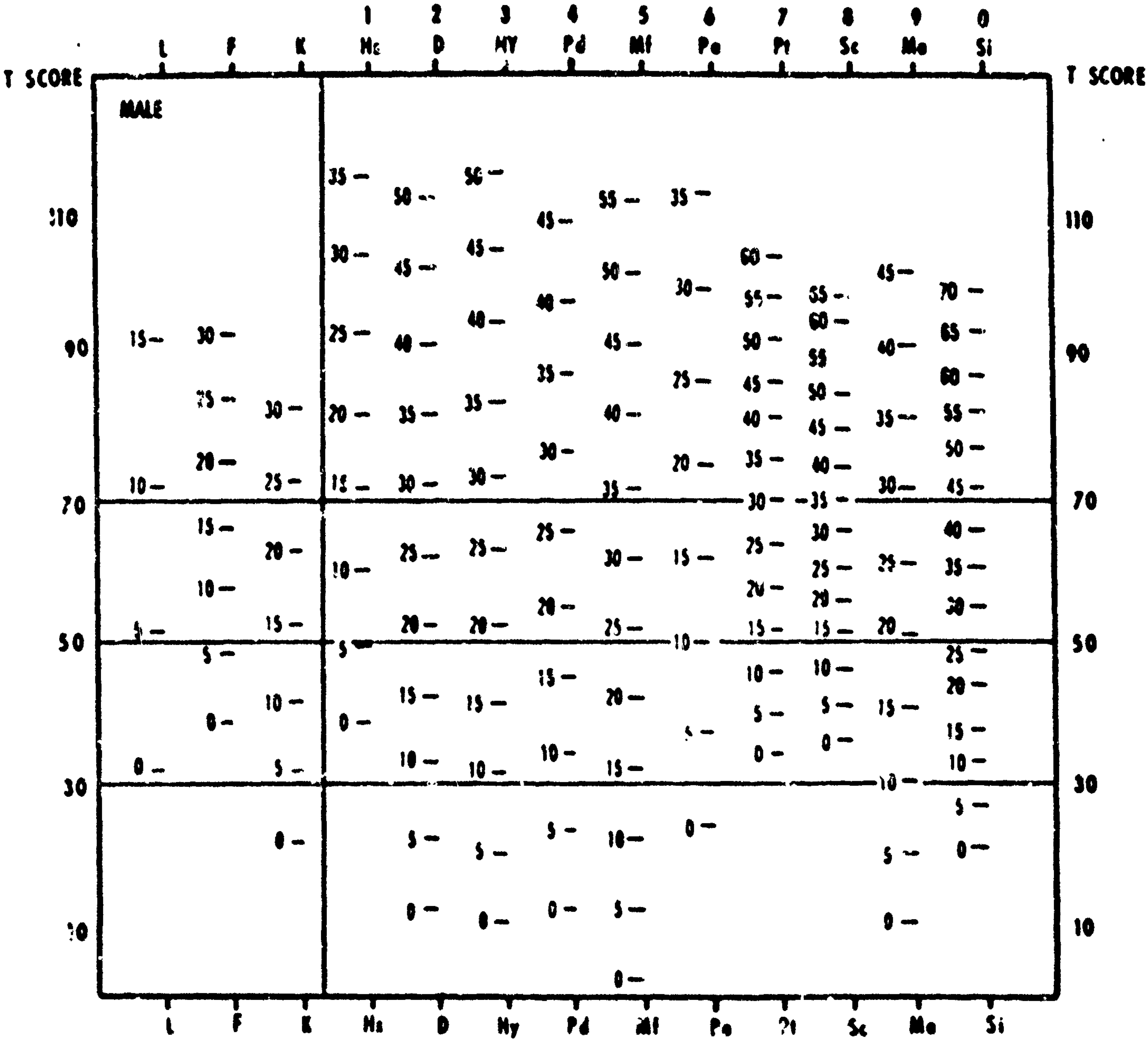
RAW SCORE

NAME SEX SSN

ILLUSTRATION 3

MALE

T SCORE CONVERSIONS FOR MMPI SCALES WITHOUT K CORRECTIONS.
BASED ON 1152 MALE USAF BASIC TRAINEES, 1975



RAW SCORE ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____

NAME SEX SSN

Table I
HOI SCORES FOR BASIC TRAINEES
Lackland Air Force Base, Texas
June 1 to November 7, 1975

HOI, High Risk Adaption Index	Number	Cumulative Number	Cumulative Percent
12 and above	299	299	0.78
11	181	480	1.25
10	347	827	2.15
9	557	1,384	3.59
8 (cut off point)	905	2,289	5.94
7	1,619	3,908	10.14
6	1,825	5,733	14.88
5	3,353	9,086	23.58
4	4,938	14,204	36.40
3	4.742	18,766	46.71
2	9,118	27,884	72.37
1	4,952	32,836	85.22
0	5,693	38,529	100.00

Table II.

EDUCATION LEVELS OF TRAINEES, LACKLAND MMPI SAMPLE
AUGUST TO OCTOBER, 1975

GRADE COMPLETED	MALE		FEMALE	
	FREQUENCY	PERCENT	FREQUENCY	PERCENT
8	0	0	1	0.1
9	3	.3	0	0
10	9	.8	12	1.5
11	26	2.3	10	1.2
12	971	84.3	638	78.3
13	82	7.1	78	9.7
14	35	3.0	38	4.7
15	8	.7	9	1.1
16	17	1.5	17	2.1
Post Grad.	1	.1	2	.2
TOTAL	1152		805	

Table III.

ETHNIC IDENTIFICATION OF TRAINEES, LACKLAND MMPI SAMPLE
AUGUST TO OCTOBER, 1975

	MALE		FEMALE	
	FREQUENCY	PERCENT	FREQUENCY	PERCENT
ANGLO-AMERICANS	999	86.7	656	81.5
BLACK-AMERICANS	104	9.0	116	14.4
HISPANIC-AMERICANS	33	2.9	2	0.2
ORIENTAL	12	1.0	22	2.7
OTHER/UNKNOWN	4	.3	9	1.1
TOTAL	1152		805	

Table IV.
AGE OF TRAINEES, LACKLAND MMPI SAMPLE
AUGUST TO OCTOBER, 1975

AGE	MALE		FEMALE	
	FREQUENCY	PERCENT	FREQUENCY	PERCENT
17	131	15.7	88	10.9
18	515	44.7	272	33.8
19	201	17.4	128	15.9
20	102	8.9	82	10.0
21	68	5.9	55	6.8
22	33	2.8	35	4.3
23	28	2.5	41	5.1
24	16	1.4	25	3.1
25	8	.7	20	2.5
26	1	.0	21	2.6
27	0	.0	15	1.8
28-34	1	.0	23	2.9
TOTAL	1152		805	

Table V.
MMPI MEANS AND STANDARD DEVIATIONS
OF BASIC TRAINEES, LACKLAND AFB
AUGUST TO OCTOBER 1975

	MALES (N=1152)		FEMALES (N=805)	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.
L	4.70	2.42	4.84	2.30
F	6.36	5.49	4.66	3.77
K	14.21	4.84	14.94	4.75
1 HS	5.51	4.52	5.02	4.04
2 D	18.96	5.04	19.11	4.41
3 HY	19.02	4.71	19.86	4.58
4 PD	17.49	4.65	16.48	4.36
5 MF	24.05	5.01	36.18	4.65
6 PA	10.14	3.97	9.42	3.37
7 PT	13.28	8.48	12.29	7.20
8 SC	13.97	10.34	11.42	8.19
9 MA	19.51	4.84	18.10	4.74
0 SI	25.62	8.93	25.59	8.88

TABLE VI.

T - Conversions for Basic MMPI Scales Without K Corrections for 885 Female, USAF Basic Trainees - Lackland Air Force Base, Texas August to October, 1973

RAW SCORE	L	F	K	1 Hs	2 D	3 Hy	4 Pd	5 Mf	6 Pa	7 Pt	8 Sc	9 Ma	0 Si	RAW SCORE
78	0	0	0	0	0	0	0	0	0	0	0	0	109	78
77	0	0	0	0	0	0	0	0	0	0	0	0	108	77
76	0	0	0	0	0	0	0	0	0	0	0	0	107	76
75	0	0	0	0	0	0	0	0	0	0	0	0	106	75
74	0	0	0	0	0	0	0	0	0	0	0	0	105	74
73	0	0	0	0	0	0	0	0	0	0	0	0	103	73
72	0	0	0	0	0	0	0	0	0	0	0	0	102	72
71	0	0	0	0	0	0	0	0	0	0	0	0	101	71
70	0	0	0	0	0	0	0	0	0	0	0	0	100	70
69	0	0	0	0	0	0	0	0	0	0	120	0	99	69
68	0	0	0	0	0	0	0	0	0	0	119	0	98	68
67	0	0	0	0	0	0	0	0	0	0	118	0	97	67
66	0	0	0	0	0	0	0	0	0	0	117	0	96	66
65	0	0	0	0	0	0	0	0	0	0	115	0	94	65
64	0	0	0	0	0	0	0	0	0	0	114	0	93	64
63	0	0	0	0	0	0	0	0	0	120	113	0	92	63
62	0	0	0	0	0	0	0	0	0	119	112	0	91	62
61	0	0	0	0	0	0	0	0	0	118	111	0	90	61
60	0	0	0	0	0	0	0	0	0	116	109	0	89	60
59	0	0	0	0	0	0	0	1	0	115	108	0	88	59
58	0	0	0	0	0	0	0	3	0	113	107	0	86	58
57	0	0	0	0	0	0	0	5	0	112	106	0	85	57
56	0	0	0	0	0	0	0	7	0	111	104	0	84	56
55	0	0	0	0	0	0	0	10	0	109	103	0	83	55
54	0	0	0	0	0	0	0	12	0	108	102	0	82	54
53	0	0	0	0	0	0	0	14	0	107	101	0	81	53
52	0	0	0	0	0	120	0	16	0	105	100	0	80	52
51	0	0	0	0	0	118	0	18	0	104	98	119	79	51
50	0	0	0	0	120	116	0	20	0	102	97	117	77	50
49	0	0	0	0	118	114	0	22	0	101	96	115	76	49
48	0	0	120	0	116	111	0	25	0	100	95	113	75	48
47	0	0	117	0	113	109	120	27	0	98	93	111	74	47
46	0	0	115	0	111	107	118	29	0	97	92	109	73	46
45	0	0	113	0	109	105	115	31	0	95	91	107	72	45
44	0	0	111	0	106	103	113	33	0	94	90	105	71	44
43	0	0	109	0	104	101	111	35	0	93	89	103	70	43
42	0	0	107	0	102	98	109	37	0	91	87	100	68	42
41	0	0	105	0	100	96	106	40	0	90	86	98	67	41
40	0	0	103	0	97	94	104	42	0	88	85	96	66	40

T - Score Conversions for Basic MMPI Scales Without K Corrections for 885 Female, USAF Basic Trainees - Lackland Air Force Base, Texas August to October, 1973

RAW SCORE	L	F	K	1 Hs	2 D	3 Hy	4 Pd	5 Mf	6 Pa	7 Pt	8 Sc	9 Ma	0 Si	RAW SCORE
39	0	0	101	0	95	92	102	44	0	87	84	94	65	39
38	0	0	99	0	93	90	99	46	0	86	82	92	64	38
37	0	0	96	0	91	87	97	48	0	84	81	90	63	37
36	0	0	94	0	88	85	95	50	0	83	80	88	62	36
35	0	0	92	0	86	83	92	53	0	82	79	86	61	35
34	0	0	90	0	84	81	90	55	0	80	78	84	59	34
33	0	0	88	119	81	79	88	57	120	79	76	81	58	33
32	0	0	86	117	79	77	86	59	117	77	75	79	57	32
31	0	120	84	114	77	74	83	61	114	76	74	77	56	31
30	0	117	82	112	75	72	81	63	111	75	73	75	55	30
29	0	115	80	109	72	70	79	65	108	73	71	73	54	29
28	0	112	77	107	70	68	76	68	105	72	70	71	53	28
27	0	109	75	104	68	66	74	70	102	70	69	69	52	27
26	0	107	73	102	66	63	72	72	99	69	68	67	50	26
25	0	104	71	99	63	61	70	74	96	68	67	65	49	25
24	0	101	69	97	61	59	67	76	93	66	65	62	48	24
23	0	99	67	95	59	57	65	78	90	65	64	60	47	23
22	0	96	65	92	57	55	63	80	87	63	63	58	46	22
21	120	93	63	90	54	52	60	83	84	62	62	56	45	21
20	116	91	61	87	52	50	58	85	81	61	60	54	44	20
19	112	88	59	85	50	48	56	87	78	59	59	52	43	19
18	107	85	56	82	47	46	53	89	75	58	58	50	41	18
17	103	83	54	80	45	44	51	91	72	57	57	48	40	17
16	99	80	52	77	43	42	49	93	70	55	56	46	39	16
15	94	77	50	75	41	39	47	96	67	54	54	43	38	15
14	90	75	48	72	38	37	44	98	64	52	53	41	37	14
13	85	72	46	70	36	35	42	100	61	51	52	39	36	13
12	81	69	44	67	34	33	40	102	58	50	51	37	35	12
11	77	67	42	65	32	31	37	104	55	48	49	35	34	11
10	72	64	40	62	29	28	35	106	52	47	48	33	32	10
9	68	62	37	60	27	26	33	108	49	45	47	31	31	9
8	64	59	35	57	25	24	31	111	46	44	46	29	30	8
7	59	56	33	55	23	22	28	113	43	43	45	27	29	7
6	55	54	31	52	20	20	26	115	40	41	43	24	28	6
5	51	51	29	50	18	18	24	117	37	40	42	22	27	5
4	46	48	27	47	16	15	21	119	34	38	41	20	26	4
3	42	46	25	45	13	13	19	0	31	37	40	18	25	3
2	38	43	23	43	11	11	17	0	28	36	38	16	23	2
1	33	40	21	40	9	9	14	0	25	34	37	14	22	1
0	29	38	19	38	7	7	12	0	22	33	36	12	21	0

TABLE VII.

T - Score Conversions for Basic MMPI Scales without K Corrections for 1152 Male USAF Basic Trainees, Lackland Air Force Base, Texas. August to October 1973

Raw Score	L	F	K	1 Hs	2 D	3 Hy	4 Pd	5 Mf	6 Pa	7 Pt	8 Sc	9 Ma	0 Si	Raw Score
79	0	0	0	0	0	0	0	0	0	0	0	0	0	79
78	0	0	0	0	0	0	0	0	0	0	112	0	109	78
77	0	0	0	0	0	0	0	0	0	0	111	0	108	77
76	0	0	0	0	0	0	0	0	0	0	110	0	106	76
75	0	0	0	0	0	0	0	0	0	0	109	0	105	75
74	0	0	0	0	0	0	0	0	0	0	108	0	104	74
73	0	0	0	0	0	0	0	0	0	120	107	0	103	73
72	0	0	0	0	0	0	0	0	0	119	106	0	102	72
71	0	0	0	0	0	0	0	0	0	118	105	0	101	71
70	0	0	0	0	0	0	0	0	0	117	104	0	100	70
69	0	0	0	0	0	0	0	0	0	116	103	0	99	69
68	0	0	0	0	0	0	0	0	0	115	102	0	97	68
67	0	0	0	0	0	0	0	0	0	113	101	0	96	67
66	0	0	0	0	0	0	0	0	0	112	100	0	95	66
65	0	0	0	0	0	0	0	0	0	111	99	0	94	65
64	0	0	0	0	0	0	0	0	0	110	98	0	93	64
63	0	0	0	0	0	0	0	0	0	109	97	0	92	63
62	0	0	0	0	0	0	0	0	0	107	96	0	91	62
61	0	0	0	0	0	0	0	0	0	106	95	0	90	61
60	0	0	0	0	0	0	0	0	0	105	94	0	88	60
59	0	0	0	0	0	0	0	120	0	104	94	0	87	59
58	0	0	0	0	0	0	0	118	0	103	93	0	86	58
57	0	0	0	0	0	0	0	116	0	102	92	0	85	57
56	0	0	0	0	0	0	0	114	0	100	91	0	84	56
55	0	0	0	0	0	0	0	112	0	99	90	0	83	55
54	0	0	0	0	120	0	0	110	0	98	89	0	82	54
53	0	0	0	0	118	0	0	108	0	97	88	119	81	53
52	0	0	0	0	116	120	0	106	0	96	87	117	80	52
51	0	0	0	0	114	118	0	104	0	94	86	115	78	51
50	0	0	0	0	112	116	120	102	0	93	85	113	77	50
49	0	0	0	0	110	114	118	100	0	92	84	111	76	49
48	0	0	120	0	108	112	116	98	0	91	83	109	75	48
47	0	0	118	0	106	109	113	96	0	90	82	107	74	47
46	0	0	116	0	104	107	111	94	0	89	81	105	73	46
45	0	120	114	0	102	105	109	92	0	87	80	103	72	45
44	0	119	112	0	100	103	107	90	0	86	79	101	71	44
43	0	117	109	0	98	101	105	88	0	85	78	99	69	43
42	0	115	107	0	96	99	103	86	0	84	77	96	68	42
41	0	113	105	0	94	97	101	84	0	83	76	94	67	41
40	0	111	103	0	92	95	98	82	0	82	75	92	66	40

T - Score Conversions for Basic MMPI Scales without K Corrections for 1152 Male USAF Basic Trainees, Lackland Air Force Base, Texas. August to October 1973

Raw Score	L	F	K	1 Hs	2 D	3 Hy	4 Pd	5 Mf	6 Pa	7 Pt	8 Sc	9 Ma	0 Si	Raw Score
39	0	109	101	0	90	92	96	80	0	80	74	90	65	39
38	0	108	99	0	88	90	94	78	120	79	73	88	64	38
37	0	106	97	120	86	88	92	76	118	78	72	86	63	37
36	0	104	95	117	84	86	90	74	115	77	71	84	62	36
35	0	102	93	115	82	84	88	72	113	76	70	82	61	35
34	0	100	91	113	80	82	86	70	110	74	69	80	59	34
33	0	99	89	111	78	80	83	68	108	73	68	78	58	33
32	0	97	87	109	76	78	81	66	105	72	67	76	57	32
31	0	95	85	106	74	75	79	64	103	71	66	74	56	31
30	0	93	83	104	72	73	77	62	100	70	66	72	55	30
29	0	91	81	102	70	71	75	60	98	69	65	70	54	29
28	0	89	78	100	68	69	73	58	95	67	64	68	53	28
27	0	88	76	98	66	67	70	56	92	66	63	65	52	27
26	0	86	74	95	64	65	68	54	90	65	62	63	50	26
25	0	84	72	93	62	63	66	52	87	64	61	61	49	25
24	0	82	70	91	60	61	64	50	85	63	60	59	48	24
23	0	80	68	89	58	58	62	48	82	61	59	57	47	23
22	0	78	66	86	56	56	60	46	80	60	58	55	46	22
21	117	77	64	84	54	54	58	44	77	59	57	53	45	21
20	113	75	62	82	52	52	55	42	75	58	56	51	44	20
19	109	73	60	80	50	50	53	40	72	57	55	49	43	19
18	105	71	58	78	48	48	51	38	70	56	54	47	42	18
17	101	69	56	75	46	46	49	36	67	54	53	45	40	17
16	97	68	54	73	44	44	47	34	65	53	52	43	39	16
15	93	66	52	71	42	41	45	32	62	52	51	41	38	15
14	88	64	50	69	40	39	42	30	60	51	50	39	37	14
13	84	62	48	67	38	37	40	28	57	50	49	37	36	13
12	80	60	45	64	36	35	38	26	55	48	48	34	35	12
11	76	58	43	62	34	33	36	24	52	47	47	32	34	11
10	72	57	41	60	32	31	34	22	50	46	46	30	33	10
9	68	55	39	58	30	29	32	20	47	45	45	28	31	9
8	64	53	37	56	28	27	30	18	45	44	44	26	30	8
7	60	51	35	53	26	24	27	16	42	43	43	24	29	7
6	55	49	33	51	24	22	25	14	40	41	42	22	28	6
5	51	48	31	49	22	20	23	12	37	40	41	20	27	5
4	47	46	29	47	20	18	21	10	35	39	40	18	26	4
3	43	44	27	44	18	16	19	8	32	38	39	16	25	3
2	39	42	25	42	16	14	17	6	29	37	38	14	24	2
1	35	40	23	40	14	12	13	4	27	36	37	12	22	1
0	31	38	21	38	12	10	12	2	24	34	36	10	21	0

TABLE VIII.

BLOOM SENTENCE COMPLETION SURVEY - ADULT

207 Basic Trainees - Lackland AFB, TX

July & August 1975

ITEM ANALYSIS OF SEVEN SCALES

SCALE NAME	SCALE SCORES	
	MEAN	SD
People	-1.0483	2.0353
Physical Self	.8019	2.0699
Family	1.1304	2.4978
Psychological Self	.6322	2.2618
Self-Directedness	1.9517	2.1440
Work	.7874	1.8581
Accomplishment	1.5459	2.1325

INTER-SCALE CORRELATIONS

	1	2	3	4	5	6	7
1 People	1.0000						
2 Physical Self	.1285	1.0000					
3 Family	.1704	.1723	1.0000				
4 Psychological Self	.2010	.3444**	.2124	1.0000			
5 Self-Directedness	.2463*	.2602*	.1022	.4299**	1.0000		
6 Work	.1902	.1938	.0768	.2422	.3491**	1.0000	
7 Accomplishment	.1285	.0387	.0084	.2407	.3966**	.1817	1.00

NOTE: *Significant at 5% level of confidence
**Significant at 1% level of confidence

TABLE IX.

BLOOM SENTENCE COMPLETION SURVEY SCORES
USAF BASIC TRAINEES, LACKLAND AFB, TX
Jan 1, 1976 to Jun 15, 1976

	ALL n = 2879		RTD n = 2213		MHC n = 666		t
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	
People	-0.5922	1.9806	-0.3954	1.9232	-1.2462	2.0274	9.61**
Physical Self	0.6544	2.2540	0.9440	2.1750	0.3078	2.2447	12.71**
Family	1.2518	2.4717	1.4867	2.3862	0.4715	2.5878	9.03**
Psychological Self	0.0327	2.1665	0.3814	2.0597	-1.1261	2.1087	16.26**
Self Directedness	2.0597	2.1344	2.4550	1.9516	0.7462	2.1893	18.10**
Work	0.9271	1.7902	1.1631	1.6627	0.1426	1.9665	12.15**
Accomplishment	2.4429	2.1486	2.8102	1.9118	1.2222	2.4223	15.52**
NET SCORES	6.7437		8.8450		-0.0976		

	Phase III-RTD n = 522		Phase III-Disch n = 114		F	t
	$\bar{x}$	sd	$\bar{x}$	sd		
People	-1.2246	2.0271	-1.3509	2.0261	0.3658	0.6056
Physical Self	-0.2264	2.2621	-0.7018	2.1148	4.2649*	2.0650*
Family	0.4982	2.5946	0.3421	2.5506	0.3054	0.5526
Psychological Self	-1.0960	2.1041	-1.2719	2.1246	0.3447	0.5871
Self Directedness	0.8442	2.1722	0.2719	2.2096	6.5201*	2.5534*
Work	0.2372	1.9439	-0.3158	2.0101	7.5603**	2.7496**
Accomplishment	1.2681	2.4629	1.0000	2.2084	1.1582	1.0723
NET SCORES	0.3008		-2.0364			

* p >.05
** p >.01

RTD - Return to Duty
MHC - Referred to Mental Hygiene Clinic

DEVELOPMENT OF

A WEIGHTED SELECTION SYSTEM

FOR

THE AFROTC PROFESSIONAL OFFICER COURSE

by

Lt Col David K. Jackson
Chief, Education Evaluation Division, AFROTC

Mr. M. Meriwether Gordon
Education Specialist, AFROTC

BACKGROUND

In April 1976, a Quality Working Group composed of representatives from the Air Staff, Air University, and AFROTC convened at Maxwell AFB, AL. Its goal was to seek alternatives for reversing a perceived downward trend in the quality of AFROTC commissionees as reflected by increasing numbers of commissionees with low AFOQT scores and by high attrition rates in the Undergraduate Flying Training Programs and in technical schools. The Quality Working Group investigated several approaches toward improving quality and placed major emphasis on the development of a weighted selection system for POC admission. Air University and AFROTC were tasked to develop a selection system for implementation on or about 3 January 1977. Initial stages in the development of the Weighted Professional Officer Course Selection System (WPSS) involved an investigation into the types of standardized tests used to evaluate potential qualification and various other factors that might influence selection of quality individuals. It was decided that Policy Capture techniques would be used to develop the Weighted POC Selection System.[1]

In October, AFROTC convened a board of nine Air Force officers and civilians to review the folders of 500 cadets entering the POC in the 76-77 school year. The board members considered approximately eighty factors on each cadet in making their selections which were then submitted to the Policy Capturing process.

The Air Force Human Resources Laboratory used regression techniques to analyze the resulting data on each of the cadets and scores resulting from the nine Policy Capture Board members. Eleven factors were found

[1]Christal, Raymond E., Selecting a Harem - and Other Applications of the Policy-Capturing Model, PRL-TR-67-1, March 1967.

to be significant predictors of the policy actually used by the Policy Capture Board. These eleven were then submitted to a process known as Heirarchical Grouping (Hier-Grp).[2]

Hierarchical Grouping (Hier-Grp) is designed to reduce a set of regression equations (also called systems or criteria) computed from proportional predictor sums of cross products matrices to a single equation. In the process, a taxonomy of regression equations results, based on the similarity between the systems. The iterative process begins with a given number of separate regression equations and, at each successive iteration, forces a compromise equation to be substituted for two of the separate systems. At each iteration, a new cluster of regression systems is formed or an existing cluster is enlarged so that, in the final step, all equations form a single broad cluster. The criterion for selecting the two equations is specified at the start of the program and is used at every step of the program.

Through the process of Hier-Grp, Human Resources Laboratory was able to develop an equation with the associated weights to be applied to each of the eleven key variables which when solved resulted in a Quality Index Score (QIS) for each applicant.

Table 1 below shows the eleven variables with their respective weights and a computation of a Quality Index Score based on a mean value for each variable and its percent of the total score.

TABLE 1

FY 79 WPSS AND THE ELEVEN VARIABLES

Variable	Mean	Wt	FY 79 WPSS Points	% Score
AFOQT-C	45.5	0.1381	6.28	(8.4)
SAT	1045.2	0.0245	25.83	(34.5)
GPA	277.3	0.1005	27.87	(37.2)
DET-CDR	3.2	1.7975	5.75	(7.7)
ASTIN[3] (Selectivity)	3.5	0.7172	2.51	(3.3)
ROTC-GPA	224.2	0.0130	2.91	(3.9)
AFOQT-Q	47.0	0.0459	2.16	(2.9)
PROGRAM	0.6	1.5837	0.95	(1.3)
TECH-MAJ	0.4	2.5949	1.04	(1.4)
NR RANKED	36.6	0.0222	0.81	(1.1)
DET-RANK	14.7	-0.0870	-1.26	(-1.7)

Quality Index Score = 74.83

[2]Bottenberg, R. A. and Christal, R. E., An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency, WADD-TN-30, AD 261 615. Lackland AFB, Texas, Personnel Laboratory, Wright Air Development Division, March 1961.

[3]Astin, Alexander W., Predicting Academic Performance in College, The Free Press, New York, 1971.

These variables were used in making selections for fall 77 admissions to the advanced program (Professional Officer Course). All selections were made or confirmed centrally at AFROTC, Maxwell AFB, although the plan was to establish a Quality Index Score cut-off point that would include about 80% of applicants who could be selected at the Detachment level. Records of the remaining 20% would be submitted to Maxwell AFB for central selection. This would insure that the best of the low-scoring group would be selected overall rather than just the best at each detachment.

RESULTS

The results of the new system to date are encouraging. Subjectively, the Detachment Commanders perceive the system as a valuable tool for assessing an applicant's potential for selection. Because of the system's objectivity and face-validity, it has been readily accepted by applicants for the program. A significant advantage of the system is that for the first time selections are made on a quantified basis that is standardized on a national level. As a result, the competition is truly national in scope. The Air University Board of Visitors placed its approval on the system in its March 77 report when it wrote: "The quality of cadets admitted to the junior and senior year programs is substantially enhanced by means of the complex instruments."[4]

The mean scores of those selected and non-selected are portrayed in Table 2 along with overall means, and means by race and sex. It may be readily noted that:

a. Black applicants are encountering difficulty in the competition, largely because of their relatively low standardized test scores.

b. The selectees are highest in Quantitative aptitude, as measured by the AFOQT. (This is probably because of the large proportion of science and technology majors in the group.)

---

[4]Air University Board of Visitors, Thirty-Third Meeting, 15 Mar 77. Minutes and Report of Chairman, (Dr. Arthur G. Hansen).

TABLE 2

ALL AFROTC APPLICANTS

	---Caucasian--		-----Black----		-----Other----		-----Total----	
	Select	Non-Sel	Select	Non-Sel	Select	Non-Sol	Select	Non-Sel
TOTAL APPLICANTS	3835	313	333	211	134	28	4302	552
TECH APPLICANTS	1729	46	103	28	61	6	1893	80
2-YR APPLICANTS	1414	191	131	109	57	11	1602	311
4-YR APPLICANTS	2421	122	202	102	77	17	2700	241
MEAN SAT (CONV)	1100.0	837.2	945.4	658.1	1052.0	777.5	1086.5	765.7
MEAN AFOQT-COMP	51.27	13.79	25.25	3.70	39.07	7.29	48.87	9.60
MEAN GPA	285.16	233.80	281.92	235.49	289.84	232.82	285.06	234.39
MEAN AFOQT-QUAN	53.11	21.17	33.27	10.60	46.65	13.54	51.38	16.74
FEMALE AFROTC APPLICANTS								
TOTAL APPLICANTS	640	67	99	70	21	5	760	142
TECH APPLICANTS	150	2	17	7	4	1	171	10
2-YR APPLICANTS	352	56	53	39	11	3	416	98
4-YR APPLICANTS	288	11	46	31	10	2	344	44
MEAN SAT (CONV)	1059.9	860.4	887.1	626.7	1042.2	798.8	1036.9	743.0
MEAN AFOQT-COMP	38.90	9.43	14.55	3.07	32.19	7.40	35.54	6.23
MEAN GPA	302.02	238.33	301.88	242.63	307.48	247.40	302.15	240.77
MEAN AFOQT-QUAN	43.71	15.57	26.16	10.07	46.71	19.20	41.51	12.99

TABLE 2 (Cont'd)

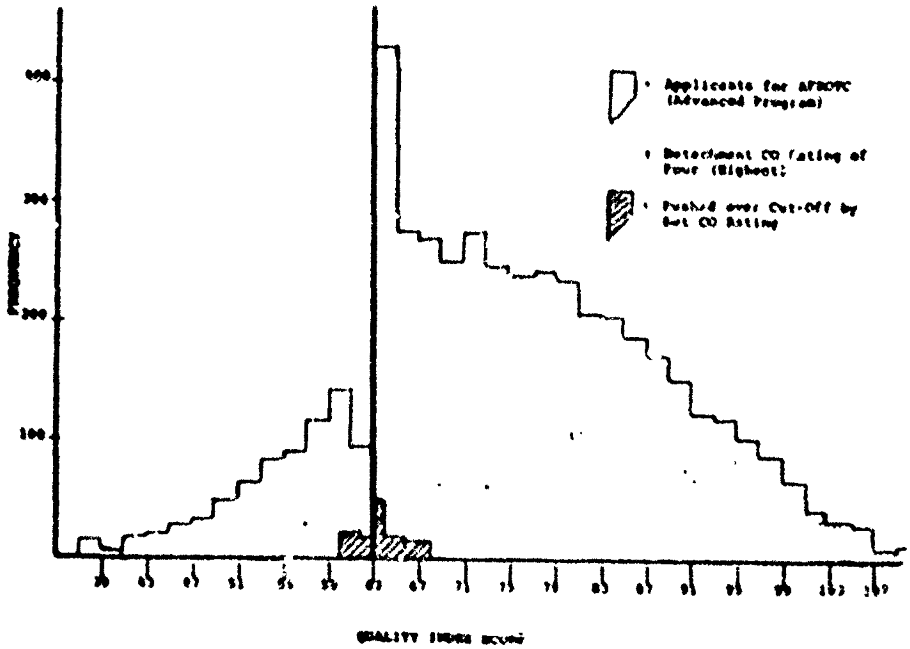
MALE AFROTC APPLICANTS

	---Caucasian--		-----Black----		-----Other----		-----Total----	
	Select	Non-Sel	Select	Non-Sel	Select	Non-Sel	Select	Non-Sel
TOTAL APPLICANTS	3195	246	234	141	113	23	3542	410
TECH APPLICANTS	1579	44	86	21	57	5	1722	70
2-YR APPLICANTS	1062	135	78	70	46	8	1186	213
4-YR APPLICANTS	2133	111	156	71	67	15	2356	197
MEAN SAT (CONV)	1108.0	830.8	970.0	673.7	1053.8	772.9	1097.2	773.6
MEAN AFOQT-COMP	53.75	14.97	29.78	4.01	40.35	7.26	51.74	10.77
MEAN GPA	281.79	232.56	273.47	231.94	286.56	229.65	281.39	232.19
MEAN AFOQT-QUAN	55.00	22.69	36.28	10.86	46.64	12.30	53.49	18.04

There is some evidence of problems with the system as indicated by the graph in Figure 1. The frequency distribution of Quality Index Scores falls into the classic bell curve, rising and falling in almost perfect, stair-case increments until it arrives at QIS 63--which was the cut-off point. There is a drastic rise in frequency just above 63 and an equally drastic drop just below it. The reasons for this asymmetry are not known; however, the cause is under continued analysis.

Figure 1

FREQUENCY OF QUALITY INDEX SCORE



The rise in overall quality as a result of the first application of the new system is clearly demonstrated in Figure 2. The rise in the Mean Officer Quality Score from 41.5 in Academic Year 76-77 to 48.8 in Academic Year 77-78 (7.3 points) was smaller than anticipated but does not tell the entire story. More significant is the redistribution of frequencies within the group resulting in a drop among those with scores below 25 from 35.8% to 21.7%. Moreover, it is now certain that those with low scores remaining have other mitigating traits.

Figure 2

UPPER AND LOWER OQC QUARTILE
% DISTRIBUTION + OVERALL MEANS

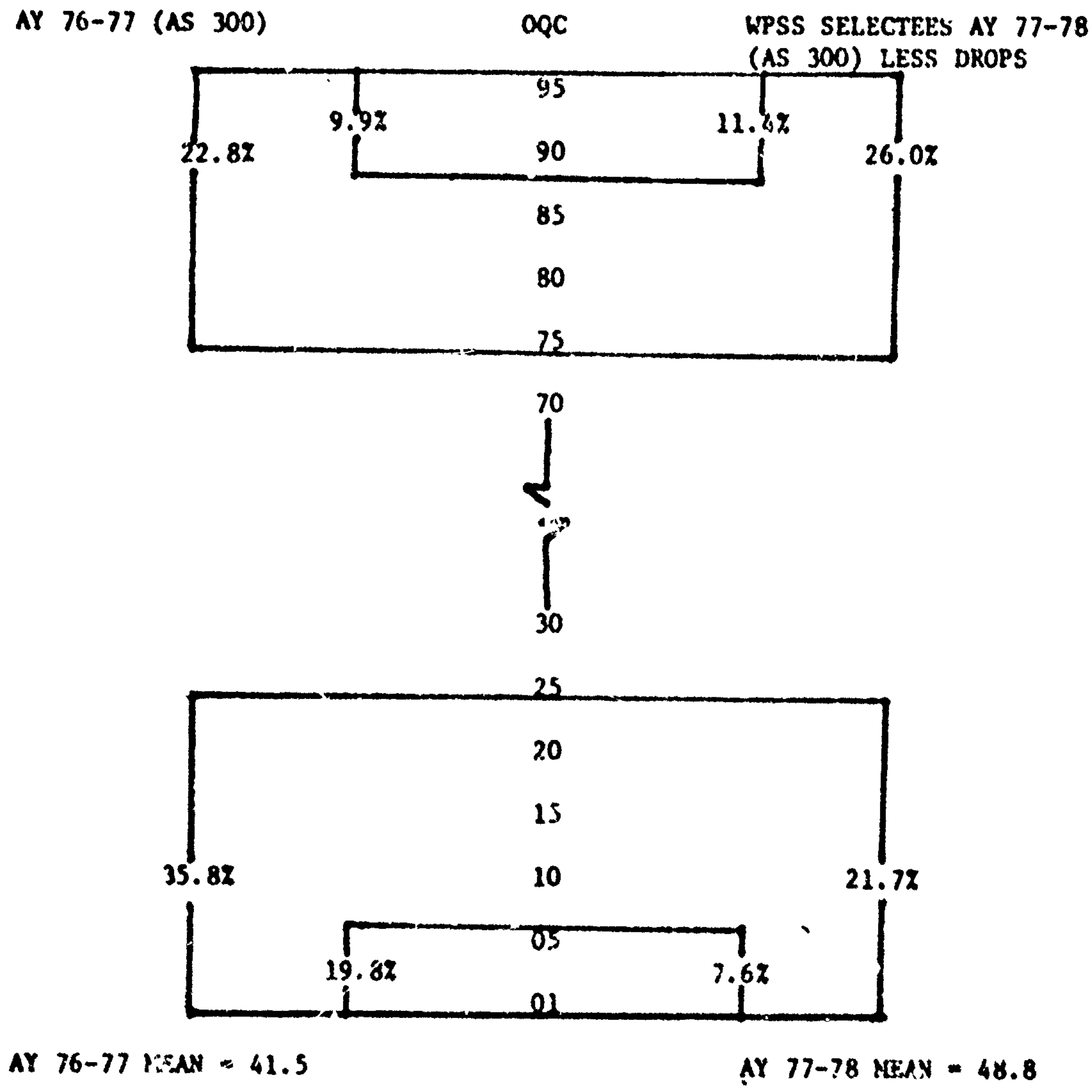


Table 3 looks at distributions of Cumulative Grade Point Averages for the WPSS selectees as compared to the AY 76/77 group. AS 300 further reveals that the distribution of higher GPA's has increased while the distribution of lower GPA's has decreased. This is clearly another indication that the WPSS is enhancing the quality of cadets who enter the POC this fall, or the AS 300 class of AY 77/78.

TABLE 3

COMPARISON OF GROUPED GPA DATA ON 76-77 ENROLLEES VS WPSS SELECTEES

	76/77 AS 300			WPSS Selectees		
	f		%	f	%	
3.75 - 4.00	107		3.5	205	4.9	
3.50 - 3.74	168	32.1%	5.4	341	8.2	40%
3.25 - 3.49	283		9.1	445	10.7	
3.00 - 3.24	436		14.1	669	16.1	
2.75 - 2.99	492		15.9	679	16.4	
2.50 - 2.74	563		18.2	680	16.4	
2.25 - 2.49	510		16.5	576	13.9	
2.00 - 2.24	420		13.6	430	10.4	
1.75 - 1.99	92		3.0	97	2.3	
1.50 - 1.74	17		.5	23	.5	
1.49	6		.2	5	.1	
	3094		100	4150	99.9	

In July 1977, a new selection board was conducted to test and validate the findings of the first board as applied through the system. The same care was taken to insure that board members were representative of all aspects of the AFROTC Program. Selections for the fall class had--for the most part--already been made, but records of a representative random sample of all applicants were furnished the board as if the selections were yet to be made. The board members were not given the Quality Index Scores but were given all the data from which the scores were computed. They were also given copies of the applicant's college transcripts, and PAS's were allowed to provide a letter in the applicant's behalf if they so desired. The findings of the board were again submitted to regression[5] and Hierarchical Grouping (Hier-Grp) procedures.

---

[5] Nie, Norman H.; Hull, C. Hadlai; Jenkins, Jean G.; Steinbrenner, Karin; Brent, Dale H., Statistical Package for the Social Sciences, 2nd Edition, McGraw-Hill Book Co., 1975.

The new weights and the influence they exerted on the total score in comparison with the old ones are listed in Table 4:

TABLE 4

FY 79 SYSTEM COMPARED WITH 77 REVALIDATION RESULTS

		-----FY 79 WPSS------			-------77 BOARD------		
VARIABLE	Mean	Wt	Points	% Score	Wt	Points	% Score
AFOQT-C	45.5	0.1381	6.28	(8.4)	0.1698	7.73	(9.3)
SAT	1054.2	0.0245	25.83	(34.5)	0.0249	26.25	(31.5)
GPA	277.3	0.1005	27.87	(37.2)	0.0894	24.79	(29.8)
DET-CDR	3.2	1.7975	5.75	(7.7)	6.2485	20.00	(24.0)
ASTIN	3.5	0.7172	2.51	(3.3)	0.7614	2.66	(3.2)
ROTC-GPA	224.2	0.0130	2.91	(3.9)	0.0046	1.03	(1.2)
AFOQT-Q	47.0	0.0459	2.16	(2.9)	0.0378	1.78	(2.1)
PROGRAM	0.6	1.5837	0.95	(1.3)	-0.2926	-0.18	(-.2)
TECH-MAJ	0.4	2.5949	1.94	(1.4)	2.9426	1.18	(1.4)
Nr RANKED	36.6	0.0222	0.81	(1.1)	0.1136	4.16	(5.0)
DET-RANK	14.7	-0.0870	-1.28	(-1.7)	-0.4152	-6.10	(-7.3)

The new weights presented some anomalies and raised some serious problems. First, the drastic change in the influence of the Detachment Commander's rating from 7.7% to 24.0% seemed to amount to a complete shift of policy from the first board to the second. Moreover, the new weight afforded the Detachment Commander's rating was largely drawn from the weights previously assigned to the SAT and GPA--measures more nearly valid across institutional lines.

Second, the slightly negative weight assigned the "Program" variable seemed illogical even though negligible. AFROTC can hardly afford to penalize enrollees in its four-year program even by 18/100ths of a point.

For these reasons, AFROTC decided to adhere to the weights derived from the first board and to ignore the findings of the second. The only change will be to round the weights to two decimal places instead of the present four and to compute future Quality Index Scores to only two decimal places. This will considerably simplify the computation and will cause only a minimal change among those selected. The modified system to be implemented 1 November 1977 is displayed in Table 5.

TABLE 5

FY 79 FOUR AND TWO DIGIT COMPARED

VARIABLE	Mean	4 DIGITS Wt	4 DIGITS Points	4 DIGITS % Score	2 DIGIT Wt	2 DIGIT Points	2 DIGIT % Score
AFOQT-C	45.5	0.1381	6.28	(8.4)	0.14	6.37	(9.2)
SAT	1054.2	0.0245	25.83	(34.5)	0.02	21.08	(30.4)
GPA	277.3	0.1005	27.87	(37.2)	0.10	27.73	(39.9)
DET CDR	3.2	1.7975	5.75	(7.7)	1.80	5.76	(8.3)
ASTIN	3.5	0.7172	2.51	(3.3)	0.72	2.52	(3.6)
ROTC-GPA	224.2	0.0130	2.91	(3.9)	0.01	2.24	(3.2)
AFOQT-Q	47.0	0.00459	2.16	(1.9)	0.05	2.35	(3.4)
PROGRAM	0.6	1.5837	0.95	(1.3)	1.58	0.95	(1.4)
TECH-MAJ	0.4	2.5949	1.04	(1.4)	2.59	1.04	(1.5)
Nr RANKED	36.6	0.0222	0.81	(1.1)	0.02	0.73	(1.0)
DET RANK	14.7	-0.0870	-1.28	(-1.7)	-0.09	-1.32	(1.9)
		QIS =	74.83		QIS =	69.45	

AFROTC will continue to monitor and adjust the system as necessary. The preliminary indications are encouraging, however, the real value of the system will not be known for several years; when attrition rates for UFT, missile school and other tech schools can be compared with accessions of new officers entering active duty under these criteria. AFROTC is sure, however, that this system will produce a better Air Force officer. The Weighted Professional Officer Course Selection System represents an equitable and workable method of selecting future Air Force officers. The results to date indicate that improvements in quality have been made.

Military Testing Association Conference

17-21 October 1977

Fitness Assessment for Entry into the Service - A Two Edged Sword?

Dennis M. Kowal, CPT, MSC, Ph.D. and James A. Vogel, Ph.D.
US Army Research Institute of Environmental Medicine
Natick, Massachusetts 01760

The Armed Forces are considering supplementing the present entry medical examination with an evaluation of physical fitness (stamina and muscle strength). Thus, new enlistees would be required to meet a minimum standard for entry into the service plus an appropriate standard for particular job specialty assignments. This proposal is based on the growing concern for the high attrition during initial training and the inability of some enlistees to physically perform their military job specialties after assignment to their units. On the other hand, there is also a concern that additional screening at enlistment will further aggrevate the anticipated shortfall in Armed Forces enlistment requirements. In response to the above requirement, the US Army Research Institute of Environmental Medicine has developed methodology for the assessment of physical fitness (work capacity) that would be suitable for use in Armed Forces Entrance Examination Stations. We now plan to evaluate the usefulness of this fitness test battery in a pilot study which will determine its ability to predict physical performance in the service. Our ultimate objective is to improve our personal selection and classification procedures for entry into the services without sexual discrimination.

Our proposed physical fitness test battery includes the following components: a) stamina (cardiopulmonary endurance or aerobic fitness); b) muscle strength of upper torso, trunk and legs; c) muscle strength endurance of upper torso and trunk; and d) coordination.

The selection of these measures was based on: a) their reliability of measurement; b) face validity as components involved in task performance; c) simplicity and ease of measurement; d) minimal equipment; and e) suitability for predicting actual physical performance in the job situation.

Stamina or aerobic fitness is measured with the use of a three load stepping test to predict maximal oxygen uptake from the exercise heart rate response (1,2). The test consists of stepping at a rate of 30 steps per minute at 3 of 4 possible step heights (depending on height of subject): 4, 8, 12 and 16 inches. Stepping continues for 3 minutes

at each step and proceeds immediately to the next step without rest. Heart rate is recorded electrocardiographically with paper stick-on electrodes and a cardiotachometer with a digital meter display. The heart rates at the end of each load are applied to a nomogram to predict maximal oxygen uptake based on the near linear relationship between heart rate and oxygen uptake and the age-maximal heart rate relation (3).

A device has been designed and built after the design of Asmussen (4) and Hermansen (5) which is used to make isometric strength measures of all three muscle groups previously mentioned. The device employs cable tensiometers, a spring-tension device that indicates kilograms of force placed on it by a cable attached between it and the muscle group of the subject. For the upper torso (shoulder-arm) muscle group, the subject is placed in a sitting position secured with a lap belt, and grasps an overhead bar so that his elbow makes a 90 degree bend and the arm is parallel to the floor. The bar is connected to the cable tensiometer to record exerted force. For the leg muscle group, the subject remains in the sitting position, and his legs flexed at 90° at the knee with his feet pushing against a bar, also connected to a tensiometer. For the trunk extensor muscle group, the subject stands facing an upright bar containing a brace plate and a shoulder strap connected to a tensiometer. The subject bends back against the shoulder strap. All three muscle group strengths are recorded by having the subject exert maximally (isometrically) for 3 seconds. Three trails are performed.

Muscle strength endurance will be measured with bent-leg sit-ups. The maximal number of sit-ups that can be performed in a 60 second period will be recorded. Upper torso strength endurance (shoulders and arms) will be tested with the flex-arm hang. In this test the maximum time is recorded that the subject can hang suspended from a bar, with their chin above the bar, hands grasping the bar.

Whole body or arm-leg coordination will be evaluated with a ladder climb test. The maximum number of rungs traversed of a 10 foot vertical ladder in 20 seconds is recorded.

Additional data to be collected along with the battery are: a) body size and body fat estimation; b) physical activity history; and c) a self report on responses to demanding situations.

A pilot study to evaluate this test battery will be conducted from January to June 1978 at the Training Center at Fort Jackson, SC. The purpose of the study is to: a) evaluate the ability of the proposed physical fitness test battery to predict subsequent physical performance during basic (BT) and advanced individual training (AIT); and b) which can be used for MOS assignment purposes (job profiling). Each component of the battery will be evaluated for its predictive validity for the

criterion referenced performances. The first criterion performance will be the new Army Physical Fitness Test which is administered during BT and AIT. This is a 3 event test: 1 mile run, sit-ups, push-ups. The second criterion performance measure consists of the Army basic common soldiering tasks: 5 mile road march, dig emplacement, 50 pound-50 meter lift and carry, hand grenade throw, 75 meter crawl, 75 meter interrupted rush. Other data to be collected will include sick call incidence, recycling and drop-out information and military performance measures.

The study design includes the initial evaluation of 1600 new basic trainees on the fitness test battery during the fill week of basic training. Criterion performance measure data will be collected as these individuals are followed through basic and advanced individual training. The Army fitness test is performed at the middle and end of BT and AIT. The common soldiering tasks will be performed at the end of BT and AIT. Upon completion of the data collection, a multiple regression analysis will be performed to specify the combination of factors (items on the test battery) that best differentiate between successful vs unsuccessful individuals in BT and AIT.

REFERENCES

1. Haisman, M.F. Assessment of the exercise capacity of young men. Ergonomics 14:449-456, 1971.

2. Shephard, R.J. et al. Standardization of submaximal exercise tests. Bull. Wld. Hlth, Organ. 38:765-775, 1968.

3. Astrand, P-O, and K. Rodahl. Textbook of Work Physiology, New York: McGraw and Hill, 1970.

4. Asmussen, E. Measurement of muscular strength. Forsvarsmed 3:152, 1967.

5. Hermansen, L. Fitness, Health and Work Capacity: International Standards for Assessment. New York:MacMillan, 1974.

Enhancing Quality Control in the

Testing of Military Applicants

M. A. Fischl

Robert M. Ross

U. S. Army Research Institute for

the Behavioral and Social Sciences

ABSTRACT

This paper acknowledges that there are diverse sources of error which must be controlled in order for aptitude tests to have substantial validity, discusses many of the sources, and describes a highly cost-effective procedure for immediate verification of the veridicality of operational test scores.

INTRODUCTION

There are both lasting and temporary, general and specific, characteristics that cause the aptitude test score an individual attains to vary from his theoretical true score. For purposes of prediction in selection and classification through the use of testing, all reasons that would increase this variance over a group may be considered error.

Such semi-permanent influences as the ability to deal with instructions on tests, or general examinee strategies for answering test questions, vary widely with individuals. The services have used several means as attempts to reduce error attributable to this "test wiseness". Instructions are easy to understand and are targeted to low levels of reading ability, and sample test items and sample instructions are provided in an information pamphlet intended to familiarize everyone concerned with the nature of the test.

Temporary influences on test scores may also affect measurement. A person's physical and emotional condition, and the physical testing environment may cause variation from true scores. To reduce these temporary effects that add to measurement error, care is taken to excuse from the testing session persons who are clearly ill or excessively fatigued, or persons who are disturbing others; and there are regulations that prohibit testing long hours without breaks or testing in places without proper conditions of lighting and temperature.

Scoring and recording errors occur either as transitory human errors or, at times, as semi-permanent conditions when, for example, an undetected malfunction develops in equipment used to score tests. Generally the variety of scoring aids now used in Armed Forces Examining and Entrance Stations (AFEES), including optical scanning, provide not only error reduction, but time savings as well.

Another source of measurement error is test compromise. These measurement errors, rather than being randomly distributed, usually operate in one direction, to yield overestimates of qualifications. Although compromise probably does not affect the measurement of very large numbers of enlistees as could other measurement errors, its non-random character makes test security of great importance.

In the past, the most common means of coping with test compromise has been by use of alternate test forms. There are two different types of alternate forms, and they differ in cost of production and in the kind of protection they provide. One type uses the same items, but arranged in different sequences in different test booklets. This type remedies situations in which the compromise has taken the form of examinees being provided with a key to the correct answers, but not the content of those answers (for example: 1a, 2c, 3d, etc.). This type of compromise is believed to be relatively uncommon. The other type of alternate test form is very much more costly to produce but also very much more comprehensive in its protection. It consists of two tests with similar (but not identical) content, matched in difficulty and other statistical properties. The protection afforded is not just for cases having the key, but for cases having the full answers to one of the forms. Both of the two types of alternate test form are now used in the test quality control programs of the services.

The parallel forms approach provides reasonable protection, but at an extremely high cost of production. That approach also does not, in and of itself, identify cases of suspect scores.

The present paper describes an alternative approach to test quality control, one which involves minimum test development costs as well as minimum examining time on site.

APPROACH

The objective of this development was to provide an operational tool to detect a substantial percentage of enlistment qualification test compromise cases. The general strategy was to capitualize on what is known or can be deduced logically concerning the differential compromise vulnerability of the various parts of the battery (ASVAB), and to combine that information with known statistical relationships among the subtests as as to "flag" highly unusual score patterns for subsequent followup.

Operational experience has shown that the main target for compromise has been the AFQT portion of the test battery. AFQT has been in joint services use the longest, for some of the services AFQT is the principal selection standard, and the nature of its contents--vocabulary, arithmetic problems, and geometric figures--are generally the best known of all military tests.

Within the AFQT portion of the battery, experience has indicated that, if compromise takes place, the compromise involves the vocabulary items by far and away the most frequently of all. This is not surprising inasmuch as vocabulary words are easy to remember and to look up after the examination. The other two subtests do not lend themselves to this kind of compromise; the arithmetic problems are relatively long prose paragraphs and there is no readily available source of the right answers as there is with a dictionary, and the totally pictorial test of spatial relations is nearly impossible to compromise through memory, and again, there is no "dictionary" available.

Given (1) that Word Knowledge may be the key ASVAB subtest compromised, that (2) the other components are relatively hard to compromise, and that (3) the psychometric relationships among these subtests are stable and known: LIKELY COMPROMISE CAN BE DETECTED BY COMPARING DISCREPANCIES IN SCORE BETWEEN THE WORD KNOWLEDGE SUBTEST AND ONE OR BOTH OF THE OTHER AFQT COMPONENTS (Arithmetic Reasoning, Space Perception).

IMPLEMENTATION

The numeric values needed to begin to implement the logic of this approach were derived from a national sample of 1,000 AFEES applicants drawn in January 1976. These 1,000 cases were stratified on AFQT to conform to the standard mobilization reference population, and the statistics shown in Table 1 were obtained. As may be seen, the correlation of Word Knowledge (WK) with Space Perception (SP) is 0.43. This means that fairly sizeable score discrepancies between WK and SP can be expected just by chance. On the other hand, the correlation of WK with Arithmetic Reasoning (AR) is high enough to be usable, 0.68. As an aside, the correlation of WK with the sum of AR and SP is no higher than the WK/AR correlation and, if used operationally, obtaining the AR/SP sum would involve an extra hand computation at operational testing sites. As a result, development focused on use of the WK/AR discrepance as being technically as sound, and operationally more feasible.

The intention was to develop a procedure which would "flag", as suspicious, cases in which the WK score exceeded the AR score by more than an expected number of points. The regression line of AR on WK can predict the expected WK score from any AR score. The prediction has tolerance bounds defined by the standard error of estimate of WK on AR and the level of significance (alpha) selected. A one - tailed

TABLE 1

STATISTICAL DESCRIPTION OF AFQT SUBTESTS OF ASVAB-6

N = 1000 AFEES APPLICANTS TESTED

IN JANUARY 1976

SUBTESTS	MEAN	S.D.	CORRELATION WITH W K
WORD KNOWLEDGE (WK), 30 ITEMS	17.5	7.5	-
ARITHMETIC REASONING (AR), 20 ITEMS	11.7	4.8	0.68
SPACE PERCEPTION (SP), 20 ITEMS	10.3	4.1	0.43

STANDARD ERROR OF ESTIMATE OF WK ON AR = 5.5

alpha level of $p \leq 0.16$ (i.e., one sigma) was chosen in consideration of maximizing detectability for subsequent followup. This resulted in identifying 11 score points as the size of the WK/AR discrepancy which would "flag" unusual cases. That is, knowing that WK and AR correlate 0.68, in a fair test a very small percentage of individuals (fewer than 16%) would be expected to exhibit a WK score 11 or more points larger than their AR score. Those who do, must be considered unusual.

A group exhibiting the unusual score pattern, consists of two types of individuals, (1) those for whom the abilities measured by the WK subtest are *truly* well in excess of their abilities in the domains measured by AR, and (2) those whose WK scores are artificially inflated through some breech of test security. The next step then, is to sort these types apart.

The simplest way to sort the compromise cases from the genuine, though unusual, ones, is to administer a 10-minute retest consisting of known secure WK items, and to compare performance on the WK retest with performance on the original WK. For some of these, the original WK score will replicate, plus or minus a calculable chance effect; for some, the second WK score will be so much lower as to be virtually unexplainable through normal chance variation.

Just as the initial screen utilized the values shown in Table 1 to define the critical WK/AR difference, values in the same Table plus those in Table 2, were used to set the chance limits for the WK 1/WK 2 difference. For this step a relatively low *alpha* level ($p \leq 0.01$) was set to minimize the risk of false accusation and to identify cases virtually unexplainable by the hypothesis of chance variation. When raw scores are converted to percentages so as to control for the different test lengths, the critical (one-sided) difference of WK 1 minus WK 2 was determined to be 26 points, and individuals exhibiting a larger difference are identified as most likely having received improper pretest assistance.

EMPIRICAL TEST

In the Spring of 1976 a sample of 111 enlistees who had been tested with ASVAB-6 at AFEES was retested at the Ft. Jackson, SC, Reception Station with ACB-73. ACB-73 contains WK and AR subtests, and was the Army's basis for computing AFQT scores until replaced by ASVAB-6 and -7 in January 1976. At the time the test sample was drawn, ACB-73 was no longer operational and, hence, its WK subtest could be looked upon as completely secure.

The first step in the test was to calculate the one-sided difference of ASVAB-6 WK minus AR, and to refer it to the specified critical

TABLE 2

STATISTICAL DESCRIPTION OF WORD KNOWLEDGE SUBTEST IN ACB-73

NUMBER OF ITEMS	MEAN	S.D.	CORRELATION WITH ASVAB-6 WK
20	11.8	4.6	0.76

STANDARD ERROR OF ESTIMATE OF ASVAB-6 WK ON ACB-73 WK = 3.0

difference of 11 points. 1/

The second step was to calculate the one-sided difference of ASVAB-6 WK minus ACB-73 WK, and refer that difference to the specified critical difference of 26 percentage points. This procedure identified 11 of the 20, or 55%, as highly suspect compromise cases. These and other important relationships are summarized in Table 3. As may be seen, if 100% of the sample had been retested on the secure WK, 23 cases would have been identified as highly suspect. Retest of only 18% of this sample identified 11 of the 23, or 48%; that is, retest of fewer than 20% of the sample "caught" almost half of the compromise cases.

A final empirical test was performed to assure maximum certainty of the percentage of the input which would have to be retested under the rule of WK-AR $\geq$ 10 points. It may be recalled that 10 points implements an alpha level of a little larger than 0.16 — i.e., a little over 16% of the population "flagged" for retesting — and one sample, at Ft. Jackson, yielded 18% so "flagged." In mid-1976, another sample of AFEES data was drawn, of size 500, and the WK minus AR criterion was again applied. Results in this sample "flagged" 17% of the cases.

SUMMARY AND CONCLUSIONS

In recognition of the fact that the Word Knowledge subtest is the most vulnerable to compromise of all the tests in the selection and classification battery, a simplified procedure was developed to detect WK compromise. The procedure has two steps:

1) At the time of scoring the AFQT portion of the battery, separate those papers in which the AR raw score is less that 15, and the WK raw score is 10 or more points greater than that AR score. This step will "flag", as potentially suspect, some 15% to 20%.

2) To only those "flagged" by step one, administer a 10-minute retest consisting of a completely secure WK, convert raw scores to percentage correct if necessary, and separate those papers in which the WK retest score is at least 26 percentage points lower than the original WK score (checklist tables can easily be prepared to accomplish all conversions and all comparisons with critical differences). This combination of setps will identify, as highly suspect, approximately half of all cases of likely test compromise.

1/ Actually two modifications were made. First, for obvious administrative simplicity, the 11 point critical difference was changed to 10. Second, cognizance also had to be taken of the difference in test length between WK and AR. Thus, the full statement of the rule became that the WK score be under 15 and the difference be 10 points.

TABLE 3

RESULTS OF EMPIRICAL TEST

	"FLAGGED" BY WK-AR	PASSED BY WK-AR	TOTAL
"CLEAN"	9	79	88
HIGHLY SUSPECT	11	12	23
TOTAL	20	91	111

SUMMARY OF

QUALITY CONTROL STEPS

- WHEN SCORING AFQT PARTS, INSPECT FOR

  AR $<$ 15 AND WK - AR $\geqq$ 10

  THIS "FLAGS" 15 - 20% OF INPUT

- TO THOSE "FLAGGED", ADMINISTER

  10-MINUTE SECURE WK RETEST

- INSPECT FOR

  $\geqq$ 26 PERCENTAGE POINT DROP WK 1 - WK 2

- THIS COMBINATION OF STEPS IDENTIFIES APPROXIMATELY

  HALF OF ALL LIKELY COMPROMISE

An alternative to the two-step procedure is to administer the WK retest to everyone and apply the rule of a 26 percentage point drop. This will detect twice the number of compromise cases, but at five to seven times the cost (that is, retesting 100% of AFEES applicants instead of between 15 and 20%).

Another alternative is to enlarge the requisite WK/AR difference so as to retest 10 percent of the input. In our Ft. Jackson sample, this detected 30 percent of the likely compromise cases.

For any of these alternatives, the conclusion may be drawn that a simple and cost-beneficial procedure for enhancing quality control in the testing of military applicants has been developed.

EVALUATION OF THE SVIB FOR PREDICTING CANADIAN MILITARY COLLEGE ATTRITION AND SUCCESS

CAPTAIN Robert W. Walker

Canadian Forces Personnel Applied Research Unit
Toronto, Ontario

Introduction

Recruitment and selection by an armed force, combined with self-selection by the volunteering candidates, tend to provide applicants to the military officer profession who have attitudes congenial to the military establishment. Nonetheless, candidates disenrol and fail from military colleges in sufficient numbers to be viewed with concern by military authorities. Such drop-outs represent failures in the transition to officer status via the academies. The problem has been present throughout the existence of military colleges, however recently the attrition rate has been alarming. Radway (1971) reported that 33% of an entering class at West Point fail to graduate. USAFA reported a 25% attrition rate (Radway, 1971) and Annapolis a 33% attrition rate (Abrahams and Neumann, 1973). Formerly, the major reasons were academic, but have recently been motivational. Attrition is greatest in the first year.

In Canada, the situation is worse. Classes entering the Canadian Military College system in the mid-sixties (1964-1967) approximated 380 cadets per year, of which approximately 230, or 60%, voluntarily disenrolled, failed or were released prior to graduation (Officer Career Development Plan, 1972). In the 1972-74 period, the attrition rate for the Royal Military College was about 50%. As in the American academies, the largest number of failures and resignations occurred in the first year of sponsorship and were mainly motivational disenrollees.

Most of the cadets who leave the Canadian Military Colleges before being commissioned can be classified as voluntary disenrollees, academic failures and military failures. Conversely, within the approximate 50% of officer cadets who do graduate and are commissioned, varying degrees of military excellence are observed and assessed during the training in the military colleges.

Research was undertaken to examine the usefulness of the Strong Vocational Interest Blank (SVIB) both for reducing the attrition of military college cadets and for predicting military excellence of those cadets retained in the system.

Retention in the Military Colleges

Research on retention in the military colleges is only a small portion of the total research on retention in the military. Two recent, comprehensive reviews of the literature applicable to retention in the military were published by Culclasure (1971) and Tuttle and Hazel (1974). Culclasure referred to extensive research conducted by naval personnel researchers who evaluated the SVIB to determine its efficacy in resolving retention problems.

Abrahams, Neumann and Githens (1968a, 1968b) evaluated the success of the SVIB for selecting naval officer cadet applicants most likely to remain on active duty beyond the minimum obligatory period. Their analysis yielded an empirical retention scale that had relatively high validity, high reliability, and low fakability. Neumann and Abrahams (1971) and Abrahams and Neumann (1971) also evaluated the SVIB as a predictor of career motivation, motivational disenrolment and academic failure of naval officer cadets undergoing college educations in civilian universities. Scales for each category of disengagement were derived that were acceptable for operational use.

The major sources of research on retention in a military college are the three major American military academies. Additional research has been completed in the Canadian military colleges. Throughout, approaches and techniques vary, reflecting the independence of operations among these institutions.

Abrahams, Neumann and Dann (1969) devised a scoring key from the SVIB to differentiate motivational disenrollees and remaining U.S. Naval Academy midshipmen following the initial, non-academic, military summer training program at the academy. Item analysis revealed clusters of items, including sports, autonomy, leadership and aesthetic interests, among others, that differentiated the two groups. An empirical scale was devised to identify midshipmen most likely to voluntarily disenrol. Abrahams and Neumann (1973) subsequently validated the SVIB for predicting not only disenrolment but also military aptitude, and three separate interest scales were empirically developed to predict motivational disenrolment, academic disenrolment and military aptitude of naval academy applicants. The results of this research clearly supported the conclusions that the SVIB was a valid predictor of midshipmen success, that scores on a derived scale were significantly related to midshipmen military aptitude ratings, and that the SVIB scales significantly aided in identifying those candidates most likely to disenrol. Additional research conducted with the SVIB was completed by Abrahams, Neumann and Githens (1970), Neumann and Abrahams (1974), Wolfe (1971), Sands (1975), and Sands and McCullah (1974).

West Point researchers have investigated various facets of the cadet success and failure enigma with varied results: admissions information (Longo, 1966); cadet interests and needs (Fishburne, 1967);

high school faculty ratings (McLaughlin, 1971a, 1971b); personality differences (Cross and Cortez, 1971); vocational interests (Marron, 1971a); military attitudes (Marron, 1971b); admission criteria (Butler, 1973, and Butler and Houston, 1974); and adjustment reactions (U'Ren, 1974). USAFA has a similar record of diversity in their research, e.g., adjustment problems (Lachar, 1974). In 1974, the Air Force academy administered the SVIB to academy applicants, after which the naval academy SVIB disenrolment scale (from Abrahams and Neumann, 1973) was scored. The reported results (from personal communication) were highly statistically significant in predicting voluntary resignations.

Published research on cadet attrition in the Canadian military colleges has originated almost totally from Royal Military College: high school academic marks (Carpenter, 1969); personality factors (Bain, 1972, 1973; Carpenter, 1973, 1974); biographical questionnaire (Shields, 1973). The conclusions were that the prediction of either distinction or attrition was an elusive undertaking.

This overview of the literature on military college cadet retention, excellence, and disenrolment indicates that the area has been investigated, but without great success. The single psychometric instrument apparently most successfully employed in specific, military environments is the SVIB. Scales derived from its item bank have predicted retention in the military college.

Method

The two aims of this study were:

(1) To develop empirical scales from the items of the SVIB to determine their potential for predicting vocational stability (attrition or retention) and vocational performance (mediocrity or excellence) in a military college.

(2) To test the validity of these empirical scales through a cross-validation study on a new sample.

Samples

Two samples of cadets were used, one for validation procedures and one for cross-validation. The first sample consisted of 400 cadets who entered Royal Roads Military College (RRMC), Victoria, B.C., Canada in 1968, 1969 and 1970 with the expectation of completing two years at RRMC plus two years at Royal Military College (RMC), Kingston,

Ontario, Canada in order to graduate from RMC in 1972, 1973, and 1974 respectively. The second sample was 266 cadets who entered RRMC in 1971 and 1972, expecting to graduate from RMC in 1975 and 1976 respectively. All cadets were administered the SVIB on entry.

The Instrument

The Strong Vocational Interest Blank is an empirically derived inventory developed by E.K. Strong shortly after World War I. In its present form (Campbell, 1969), the SVIB for Men consists of 399 items drawn from several areas of life -- occupations, school subjects, amusements, kinds of people, work situations, etc. There are 54 occupational scales, 22 Basic Interest scales and 8 non-occupational scales in the SVIB.

Validation Procedure

In order to develop empirical scales that would discriminate between groups, it was first necessary to define reference and criterion groups. For the final status or disposition categories, the task was not difficult. The 400 cadets in the validation sample were classified as graduated on schedule (N=167), voluntarily disenrolled (N=104), academically failed (N=37), and militarily failed (N=38). Fifty-four (13.5%) of the sample did not fit into these categories and were removed from the analysis. In the construction of the scales, the graduates constituted the reference group and voluntary disenrollees or academic failures or military failures constituted the criterion groups, each in separate analyses.

For the scales constructed to discriminate excellence or mediocrity in vocational performance, only those cadets reaching fourth year were used in the analyses. Performance was assessed by academic grades, military grades and cadet officer appointment level. The academic mediocrity/excellence categories consisted of a reference group with third class, pass or borderline academic grades (N=82) and a criterion group with second and first class grades (N=88). The military grades, determined by commissioned officers whose primary responsibility was the care, management and assessment of squadrons of cadets, permitted the classification of cadets into mediocrity and excellence categories. These consisted of a reference group with military grades of C-, C or C+ (N=118) and a criterion group with grades of A or B (N=45).

The cadet wing is partially self-regulated by a cadre of fourth year cadets assigned to "cadet officer appointment" positions. The cadet holding the highest position is called the Cadet Wing Commander (air force rank terminology) and wears five stripes or bars on his tunic, hence holds a 5-bar position. Bar positions exist at the 5, 4, 3 and 2-bar levels. Fourth year cadets not selected for cadet officer appointments receive no bars. Therefore, for the cadet officer appointment level or "bars" categories, the reference group consisted of fourth year cadets

assigned 0 or 2 bars (N=98) and the criterion group consisted of fourth year cadets assigned 3 or 4 bars (N=72). No cadets in the sample received 5 bars.

Considerable controversy exists in the psychometric literature as to the appropriate scale length for scales intended for discriminating between groups. Too few items (e.g., less than 40) reduce reliability while too many items emphasize chance differences or create unnecessary redundancy (Campbell, 1971). Campbell suggested a scale length of not more than 100 items and preferably not less than 60, but if necessary, a minimum of 40. In this project, to ensure that sufficient items were used while at the same time selecting only items that discriminated between groups with percentage differences of acceptable magnitude ($\geq$ 10%, Abrahams and Neumann, 1973), it was decided that two scales for each of the six criteria would be constructed — a 50-item and a 75-item scale. Therefore, 12 empirical scales were constructed.

The precise method for selecting and then weighting discriminating items has varied with new developments and new instruments in psychometrics. In an exceptionally thorough investigation, Sands (1975) developed and evaluated many item response weighting procedures. For the problems addressed in his study, almost all of the different weighting methods had essentially the same ability to differentiate between groups. Sands suggested, therefore, that the simplest procedures for weighting continue to be used.

In this project, items showing differences in percent responses of $\geq$ 10% between the reference and criterion groups were identified and given unit weights in the appropriate direction. The items were then dimensionalized by weighting the opposite end of the items in the reverse direction. The Indifferent response was weighted in the appropriate direction (+1 or -1) if the percentage difference between the criterion and reference group was about 10% or more. Otherwise, it was given a weight of 0. Constants of 50 and 75 were added to the 50-item and 75-item scales respectively to convert all scores to positive values in scale ranges from 0 - 100 and 0 - 150.

Validation of the scales was investigated by applying the scales to the scale development sample. A comparison of reference group and criterion group scores for each scale yielded biserial correlations which indicated the scale's validity for predicting the particular criterion being assessed. As well, expectancy tables were constructed to demonstrate graphically the magnitude of relationships between the scale scores and criterion variables.

Cross-validation Procedure

This second phase constituted a cross-validation study on a new sample to test the validity of any relationships found above (i.e., validity generalization). The comparison of predicted and actual outcomes

in this new sample determined whether the empirical scale development method could be used to reduce attrition losses in the military college system.

Results

Validation Phase

Empirical Scales. Table 1 provides descriptive statistics for the 12 empirical scales constructed. Same-named scales (in the order listed in Table 1) of 50-item and 75-item lengths correlated .950, .950, .916, .951, .969 and .952, all $p < .001$, indicating that considerable redundancy and little benefit resulted from the difference in scale lengths.

Discriminating Capacity. The ability of the scales to discriminate between the criterion groups and the reference groups was assessed by calculating biserial correlations as estimates of the relationship between the scale scores, distributed continuously, and the criterion and reference groups, which together constituted a dichotomous variable. Biserial correlations were used instead of point biserial correlations (Abrahams and Neumann, 1973) since the variables underlying the dichotomies were assumed to be continuous and normal (e.g., a continuum of predisposition to voluntarily disenrol).

For all 12 scales, the biserial correlations were significantly different from 0. Values ranged from .722 to .935, which indicated that the scale scores were distributed significantly differently between criterion and reference groups. These results were not surprising, however, since the scales were being applied to the same sample used to develop them. This procedure artificially inflated the statistics, reflecting a relationship that could be expected to shrink in a cross-validation sample.

The expectancy table in Figure 1 was prepared as an example to demonstrate graphically the magnitude of the relationships between the scales and the criteria. The expectancy table was constructed by dividing the scale score distributions into quintiles and assessing these quintiles as to their proportion of cases from the reference group and the criterion group. Figure 1 depicts the increasing proportion of voluntary disenrollees found in each quintile of the scale score distribution. Following this example further, for the 75-item voluntary disenrolment scale, voluntary disenrollees in the lowest through highest quintiles constituted approximately 3%, 10%, 24%, 31% and 65% of the quintiles' population. Only one of the twelve expectancy tables was presented here since the trends were very similar across all scales.

TABLE 1

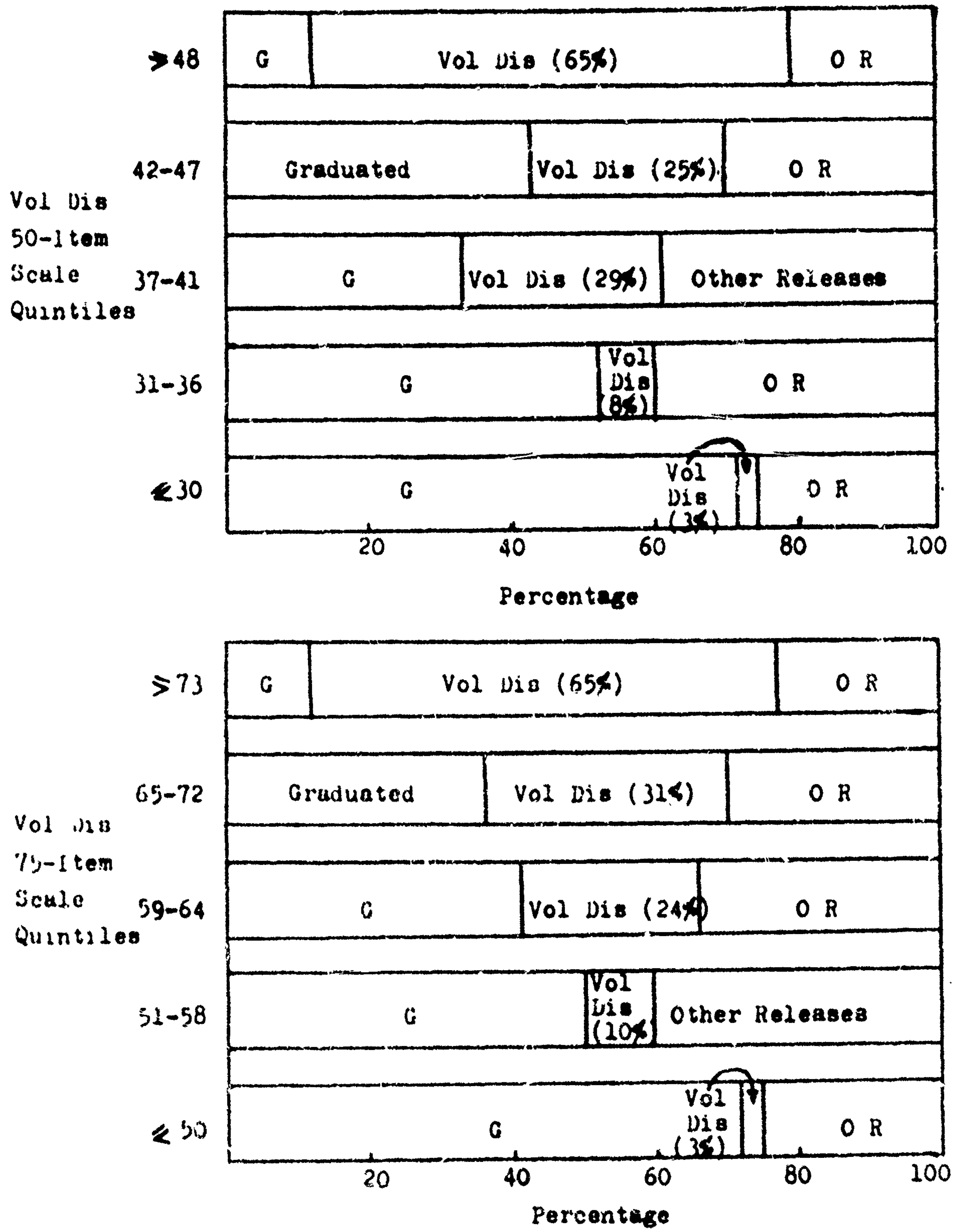
Descriptive Statistics of 12 Empirical Scales: Validation

Scale	N	Mean	Standard Deviation	Range[1]	Minimum	Maximum
50-Item Scales						
Voluntary Disenrolment	400	39.7	10.0	58	9	67
Academic Failure	400	52.5	9.4	53	24	77
Military Failure	400	47.0	8.8	53	24	77
Academic Performance	167	46.4	10.5	51	19	70
Military Performance	167	46.9	11.0	48	25	73
Bars	167	47.4	10.4	49	18	67
75-Item Scales						
Voluntary Disenrolment	400	61.8	12.3	75	27	102
Academic Failure	400	77.9	11.8	70	41	111
Military Failure	400	74.8	10.8	60	47	107
Academic Performance	167	66.8	13.7	63	37	100
Military Performance	167	72.3	14.2	62	47	109
Bars	167	74.6	14.0	64	39	103

[1]Maximum possible range for 50-item scales = 0 - 100, for 75-item scales = 0 - 150.

FIGURE 1

Expectancy Tables Depicting Likelihood of Voluntary Disenrolment Associated with Scores on the Voluntary Disenrolment Empirical Scales: Validation



Cross-validation Phase

The ability of the empirical scales to discriminate between criterion and reference groups in the cross-validation sample was indicated by the magnitude of the biserial correlations. For 6 of the 12 scales, the biserial correlations were significantly different from 0 (Table 2). The two voluntary disenrolment scales possessed the highest discriminating abilities, the two bars scales the next highest. The two military performance scales possessed discriminating abilities that just reached significance. The two academic failure scales, the two military failure scales and the two academic performance scales failed to discriminate significantly between criterion and reference groups. Table 2 contains criterion and reference group means, biserial correlations and significance levels for all twelve scales. The "shrinkage" anticipated from applying the scales developed in the validation sample to the cross-validation sample did occur.

Due to several considerations not presented thoroughly in this paper (e.g., lack of discriminating ability as indicated in expectancy tables for military performance, the redundancy in scales in that military performance scales and bars scales were essentially measuring military success, the fact that the bars scales were discriminating better, etc.), the decision was made to drop the military performance scales from further consideration. This, in essence, meant that, for predicting attrition, the two voluntary disenrolment scales would be used and, for predicting military excellence among graduates, the two bars scales would be used. The expectancy tables for these scales are presented in Figures 2 and 3.

For these four remaining scales, item - scale score correlations were calculated to identify those items not contributing significantly to the scale totals. Items not significantly correlated were removed from the scales, producing new scales of shorter length (renamed VODI45, VODI55, BARS40 and BARS57 to reflect scale length). Biserial correlations calculated for these abbreviated, "purified" scales indicated that the discriminating abilities of these scales were not significantly improved.

In the expectancy tables for the revised voluntary disenrolment scales (Figure 4), all distributions in all quintiles were within a few percentage points of the same distributions before the scales were revised (Figure 2). For the revised bars scales, the distributions in the quintiles appeared slightly improved in that the extreme cases (0 bars and 4 bars) were better distributed across the quintiles. (Compare Figures 3 and 5).

Tests of internal consistency were conducted using split-half reliabilities to determine if revising the scales through removal of items not correlated with scale scores had improved the scales. For the VODI45 and VODI55 scales, correlations between test halves equalled .478 and .582 that, when corrected with the Spearman-Brown formula, equalled .647 and .735, $p < .001$, compared to .592 and .567 before

TABLE 2

Biserial Correlations in Empirical Scale Analysis; crossvalidation

Scale		Mean (Criterion Gp)	Mean (Reference Gp)	Standard Deviation	Biserial Correlation	Significance Level
Vol Dis	50	44.4	36.9	9.1	.510	.001
Vol Dis	75	67.3	58.2	11.2	.504	.001
Acad Fail	50	54.7	51.4	8.3	.223	n.s.
Acad Fail	75	81.2	77.5	10.3	.200	n.s.
Mil Fail	50	44.8	46.6	7.6	-.126	n.s.
Mil Fail	75	72.0	74.9	9.7	-.158	n.s.
Acad Perf	50	45.7	44.2	9.2	.098	n.s.
Acad Perf	75	65.8	63.5	11.5	.122	n.s.
Mil Perf	50	48.6	45.2	8.4	.250	.05
Mil Perf	75	75.1	70.1	10.7	.291	.05
Bars	50	50.5	46.3	8.2	.323	.01
Bars	75	78.1	72.7	10.8	.310	.01

FIGURE 2

Expectency Tables Depicting Likelihood of Voluntary Disenrolment Associated with Scores on the Voluntary Disenrolment Empirical Scales; Crossvalidation

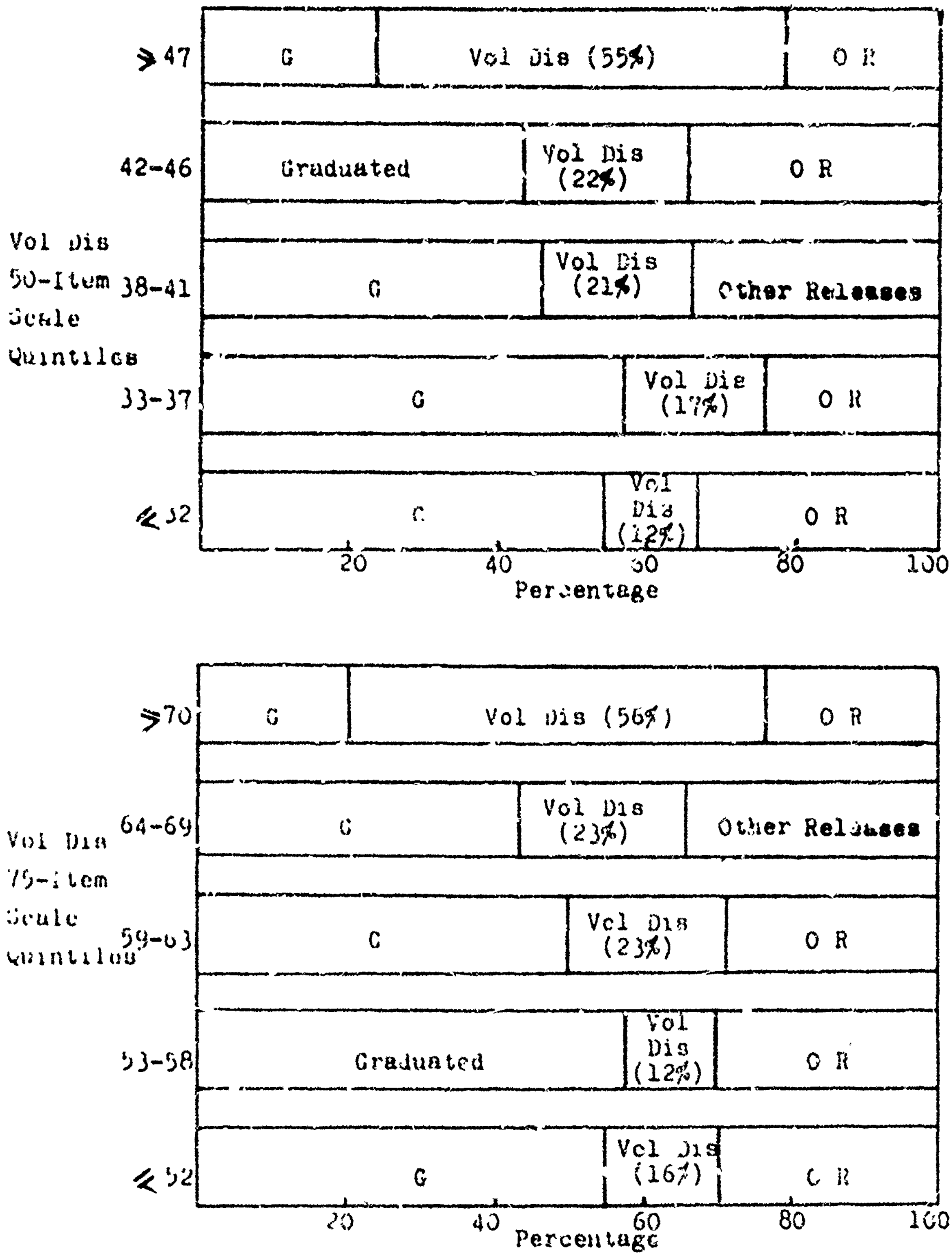


FIGURE 3

Expectancy Tables Depicting Likelihood of Bars Associated with Scores on the Bars Empirical Scales; Crossvalidation

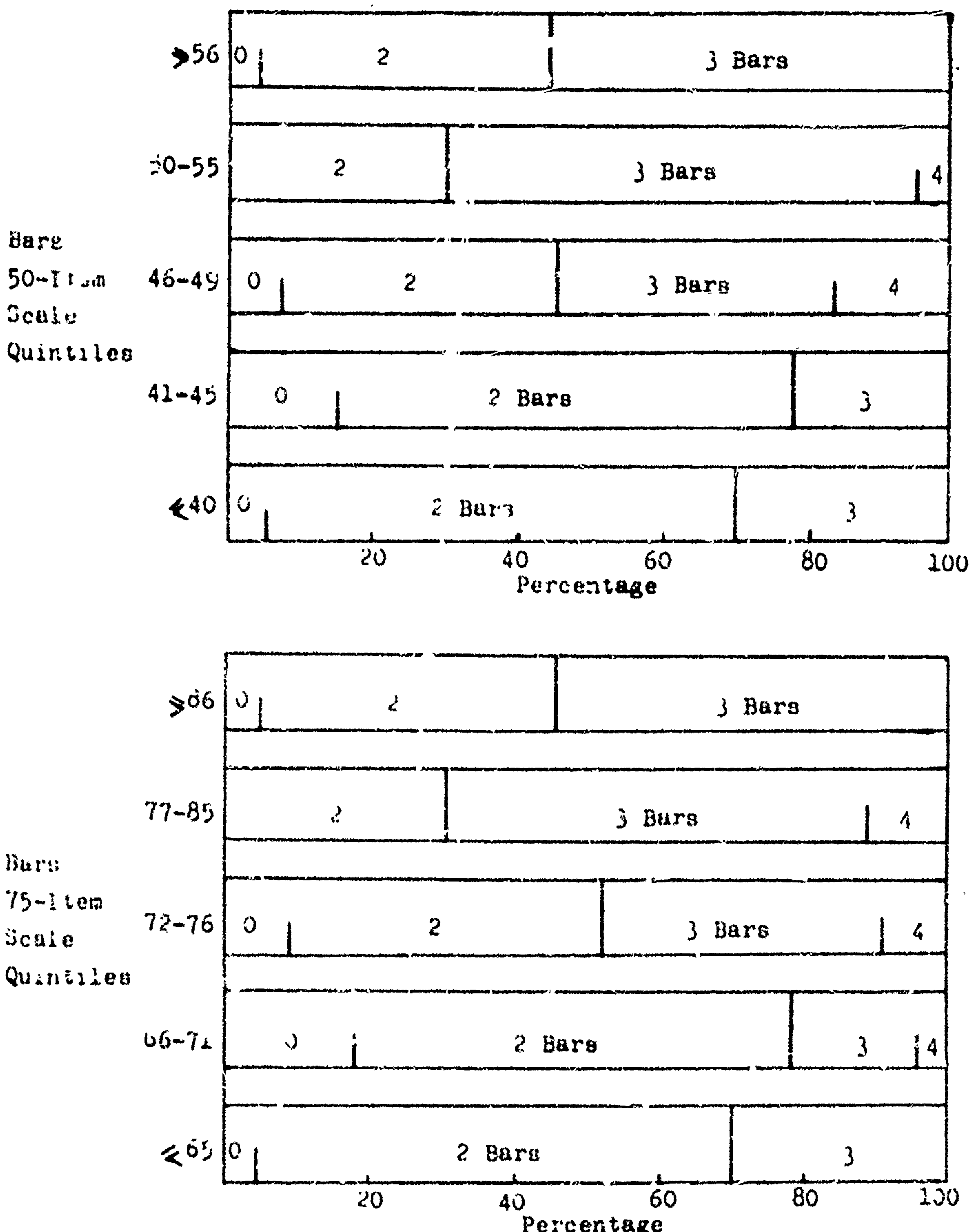


FIGURE 4

Expectancy Tables Depicting Likelihood of Voluntary Disenrolment Associated with Scales on the Revised Voluntary Disenrolment Empirical Scales; Crossvalidation

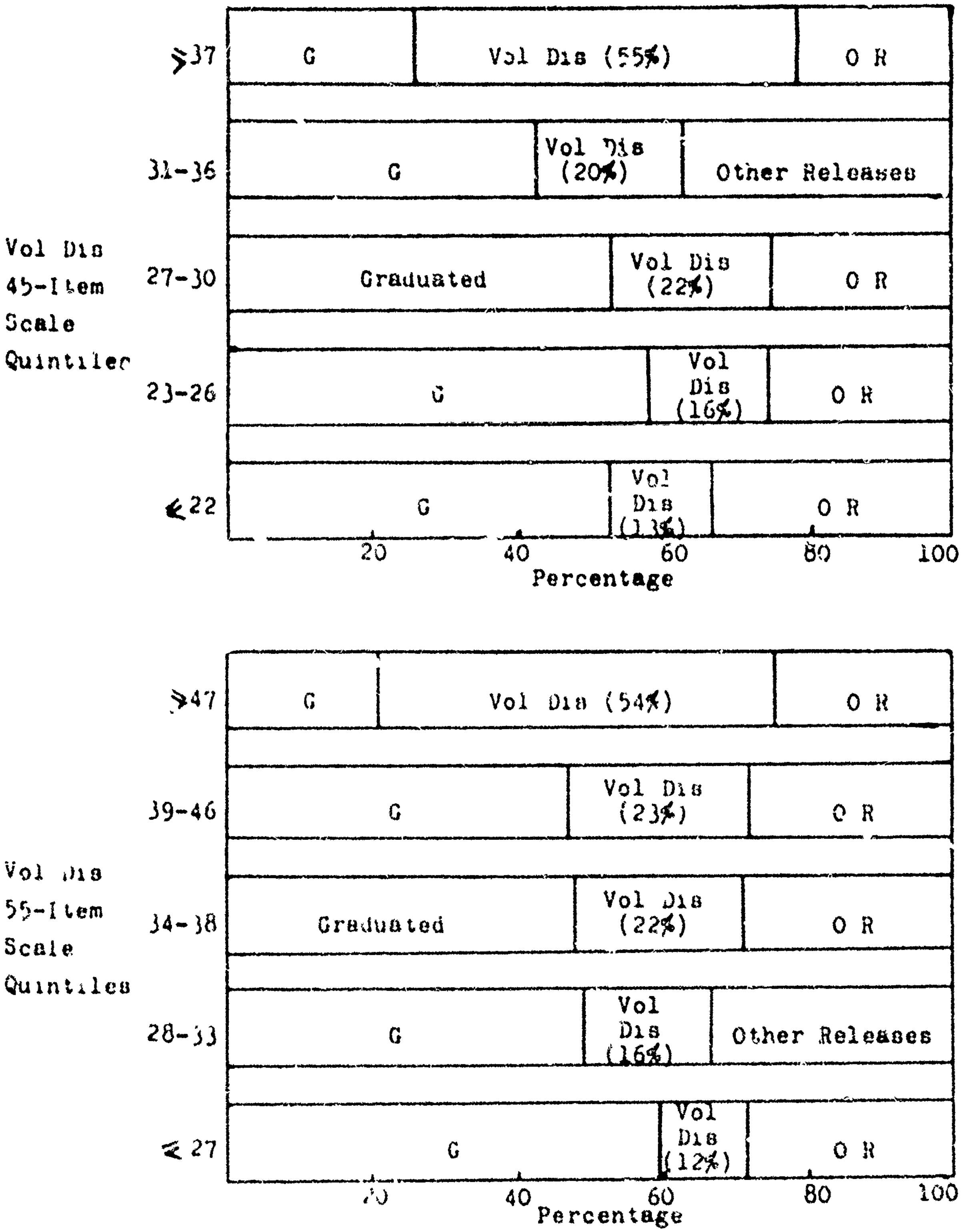
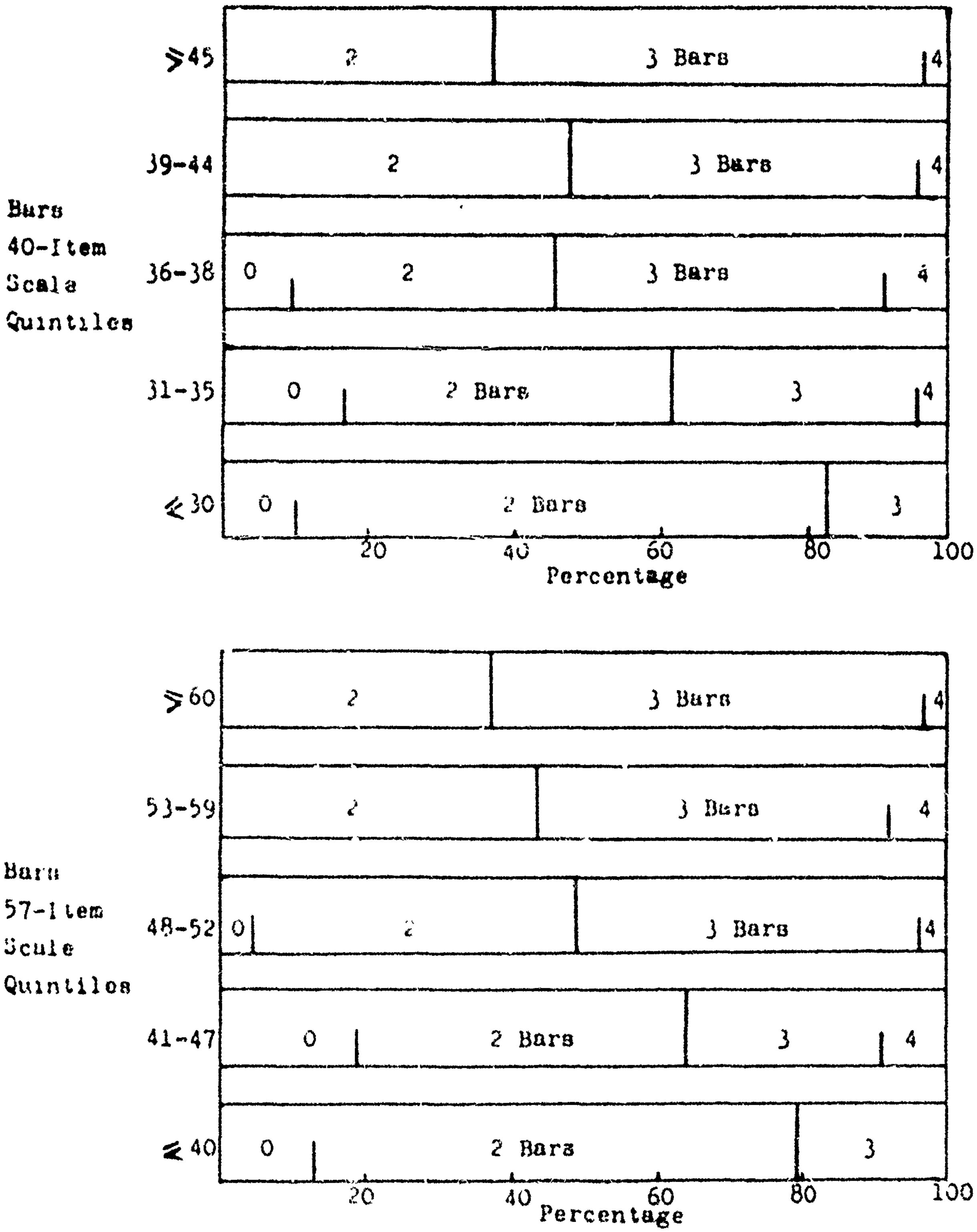


FIGURE 5

Expectancy Tables Depicting Likelihood of Bars Associated with Scores on the Revised Bars Empirical Scales; Crossvalidation



revision. It appeared to be the case that scale reliability was improved by the scale revisions, as indicated by the reliability coefficients, however, the predictive validity of the scales did not significantly change.

The potential advantage in using the scales for selection could be exemplified using the voluntary disenrolment likelihoods outlined in Figure 4. Those candidates scoring in the top quintile of the VODI55 scale were 4 1/2 times more likely to disenrol than candidates in the bottom quintile. For Figure 5, candidates scoring in the top quintile in the BARS40 scale were 3 times more likely to achieve 3-bar or 4-bar cadet appointments than those scoring in the bottom quintile.

To identify the psychological traits underlying the differences in scale scores between criterion and reference groups, factor analyses of the scale items were conducted. A principal factors with Varimax orthogonal rotations factor analytic technique (Nie, et al., 1975) was carried out with the items in the VODI and BARS scales. The factor accounting for the largest percentage (29.6%) of variance in the voluntary disenrolment scale contained 8 items concerned with artistic/aesthetic interests (e.g., art museum director) which were endorsed as "liked" by voluntary disenrollees. The second factor, containing 8 items representing engineering/mechanical interests (e.g., electronics equipment designer) and the third factor, which included military activities (e.g., military drill), both consisted of items that were responded to in the negative by voluntary disenrollees. The fourth factor contained items reflecting acceptance of nonconformity in others (e.g., like, as people, daydreamers, beachcombers) to which voluntary disenrollees responded positively. The fifth factor included items representing military types (e.g., General of the Army) to which voluntary disenrollees responded negatively. Total variance accounted for by the first five factors was 85.4%. In all, the candidates opting for early disenrolment appeared to have interests that were more artistic or aesthetic, less mechanical, less oriented toward military activities and military people, and more liberal toward nonconformity.

The five clusters of items differentiating militarily excellent cadets from mediocre cadets, according to the bars scales, were all endorsed more positively by excellent cadets than mediocre cadets. The first cluster included social-artistic interests (e.g., dramatics, psychologist) and the second cluster represented athletic and recreational leadership (e.g., athletic director). The third cluster involved social responsibilities (e.g., starting an activity) while the fourth cluster consisted of items on teaching and directing others (e.g., teach children, scout leader). The final cluster involved items reflecting enterprising endeavours (e.g., stockbroker, corporation lawyer). Collectively, these factors accounted for 92.9% of the variance, and indicated that cadets who excelled in the military college system, generally, had more interests in aesthetic, athletic, social

and enterprising activities, and like to teach and direct others. The converse appeared to be that mediocre cadets had interests in mechanical but not aesthetic areas, were less interested in athletics, were more reserved, were not enterprising and preferred not to teach or direct others. Hindsight allowed the observation that these factors basically identified the characteristics desired in a strong leader: an artistic, creative, flexible orientation; physically active and involved; extroverted and capable of infusing life into an organization; interested in teaching or instructing others; and possessing an enterprising manner and ability to influence others.

Conclusions

The empirical scales, developed to predict attrition or retention, and excellence or mediocrity for retained cadets, showed a strong capacity to predict. The scales designed to predict probabilities for voluntary disenrolment did successfully differentiate between cadets who did voluntarily resign and cadets who stayed and graduated. As well, separate scales differentiated between mediocrity and excellence in cadet performance as indicated by the level of cadet officer appointment (bars) assigned in fourth year. This research supports the conclusion that the SVIB is a valid predictor of military college attrition or of retention with excellent performance. On the basis of these results, it is recommended that the SVIB be further investigated for its full potential to predict early disenrolment and military excellence!

References

Abrahams, N.M., and Neumann, I. The assessment of career motivation among NROTC applicants with the Strong Vocational Interest Blank. Research Report SRR 72-9. San Diego: Naval Personnel and Training Research Laboratory, October, 1971.

Abrahams, N.M., and Neumann, I. The validation of the Strong Vocational Interest Blank for predicting Naval Academy disenrollment and military aptitude. Technical Bulletin STB 73-3. San Diego: Naval Personnel and Training Research Laboratory, February, 1973.

Abrahams, N.M., Neumann, I., and Dann, J.E. Use of the Strong Vocational Interest Blank in identifying Naval Academy early motivational disenrollees. Research Report SRR 70-4. San Diego: U. S. Naval Personnel and Training Research Laboratory, October, 1969.

Abrahams, N.M., Neumann, I., and Githens, W.H. The Strong Vocational Interest Blank in predicting NROTC officer retention: Part I, validity and reliability. Technical Bulletin STB 68-7. San Diego: U. S. Naval Personnel Research Activity, January, 1968(a)

Abrahams, N.M., Neumann, I., and Githens, W.H. The Strong Vocational Interest Blank in predicting NROTC officer retention: Part II, fakability. Technical Bulletin STB 68-9. San Diego: U.S. Naval Personnel Research Activity, February, 1968. (b)

Abrahams, N.M., Neumann, I., and Githens, W.H. An assessment of faking on the Strong Vocational Interest Blank under actual selection conditions. Technical Bulletin STB 70-6. San Diego: Naval Personnel and Training Research Laboratory, March, 1970.

Bain, E.H. Personality, leadership and motivational characteristics of ROTP resignees from RMC. MLM Research Report 72-8. Kingston, Ontario: Royal Military College of Canada, October, 1972.

Bain, E.H. The socialization process of ROTP cadets at RMC. MLM Research Report 73-3. Kingston, Ontario: Royal Military College of Canada, July, 1973.

Butler, R.P. Predictors of academic performance. IA4.01-74-003. West Point, N.Y.: Office of the Director of Institutional Research, United States Military Academy, August, 1973.

Butler, R.P., and Houston, J.W. Attrition and admission scores, class of 1975. 2A1.00-74-019. West Point, N. Y.: Office of the Director of Institutional Research, United States Military Academy, January, 1974.

Campbell, D.P. Manual (1969 Supplement) for Strong Vocational Interest Blank. Stanford, Calif.: Stanford University Press, 1969.

Campbell, D.P. Handbook for the Strong Vocational Interest Blank. Stanford, Calif.: Stanford University Press, 1971.

Carpenter, G.J. The validation of variables available for the selection of cadets to the Canadian military colleges. Paper presented at the Eleventh Annual Conference of the Military Testing Association, New York City, September, 1969.

Carpenter, G.J. Cadet retention at the Royal Military College of Canada as a function of personality, leadership performance, and biographical variables. MLM Research Report 73-4. Kingston, Ontario: Royal Military College of Canada, October, 1973.

Carpenter, G. J. The relationship between personality factors and the achievement of distinction at the Royal Military College. MLM Research Report 74-4. Kingston, Ontario: Royal Military College of Canada, August, 1974.

Cross, W.M., and Cortez, J.J. A comparative personality study of resignees at USMA. Paper presented at the 2nd Annual Symposium of Psychology in the Air Force, USAFA, Colorado Springs, Colorado, April, 1971.

Culclasure, D.F. Development of career motivational prediction and selection procedures. AFHRL-TR-71-19. Brooks Air Force Base, Texas: Personnel Division, Air Force Human Resources Laboratory, January, 1971.

Fishburne, F.J., Jr. An investigation of the use of the Activities Index and the College Characteristics Index as predictors of voluntary attrition at the U.S. Military Academy. Unpublished M. A. Thesis, Ohio State University, 1967.

Lachar, D. Prediction of early U.S. Air Force freshman cadet adaptation with the MMPI. Journal of Counseling Psychology, 1974, 21, 404-408.

Longo, A.A. Maximizing the prediction of overall academy performance. IA3.04-66-001. West Point, N.Y.: Office of Research, United States Military Academy, September, 1966.

Marron, J.E. Validation of the academic E-I, leadership E-I, and resignation E-I scores. IA3.03-72-006. West Point, N.Y.: Office of Institutional Research, United States Military Academy, July, 1971. (a)

Marron, J.E. The validity of the military attitude and association scales. IB5.05-71-006. West Point, N.Y.: Office of Institutional Research, United States Military Academy, May, 1971. (b)

McLaughlin, G.W. The use of high school faculty ratings to predict USMA fourth class performance. IA4.06-71-001. West Point, N.Y.: Office of Institutional Research, United States Military Academy, March, 1971. (a)

McLaughlin, G.W. A computer assisted selection system for USMA. IA3.08-72-012. West Point, N.Y.: Office of Institutional Research, United States Military Academy, September, 1971. (b)

Neumann, I., and Abrahams, N.M. The development of an SVIB scale for predicting disenrollment from the NROTC (Regular) program. Research Memorandum SRM 71-10. San Diego: Naval Personnel and Training Research Laboratory, February, 1971.

Neumann, I., and Abrahams, N.M. Identification of Naval Academy applicants with engineering and science interests. NPRDC-TR-75-7. San Diego: Navy Personnel Research and Development Center, October, 1974.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.H. Statistical package for the social sciences (SPSS). (2nd ed.). New York: McGraw-Hill, 1975.

Officer career development program. Ottawa: Department of National Defence, Government of Canada, 1972.

Radway, L.I. Recent trends at American service academies. In M.R Van Gils (Ed.), The perceived role of the military. Rotterdam: Rotterdam University Press, 1971.

Sands, W.A. Alternative item response weighting procedures: Development and evaluation. Paper presented at the 17th Annual Conference of the Military Testing Association, Fort Benjamin Harrison, Indiana, September, 1975.

Sands, W.A., and McCullah, R.D. Development and evaluation of Strong Vocational Interest Blank scales for academic major areas at the U.S. Naval Academy. Paper presented at the 16th Annual Conference of the Military Testing Association, Oklahoma City, Oklahoma, October, 1974.

Shields, W.S. The use of qualitative information in the prediction of college attrition. MLM Research Report 73-6. Kingston, Ontario: Royal Military College of Canada, 1973.

Tuttle, T.C., and Hazel, J.T. Review and implications of job satisfaction and work motivation theories for Air Force research. AFHRL-TR-73-56. Brooks Air Force Base, Texas: Occupational Research Division, Air Force Human Resources Laboratory, January, 1974.

U'Ren, R.C. The nature of change at West Point. Journal of the American College of Health Association, 1974, 22, 185-188.

Wolfe, J.H. Comparative cluster analysis of patterns of vocational interest. Technical Bulletin STB 72-3. San Diego: Naval Personnel and Training Research Laboratory, October, 1971.

Effects of Test Item Formats on Difficulty and Reliability

David S. Vaughan, Capt., USAF
ATC Technology Applications Center

and

William Fogarasi, SrA., USAF
USAF Occupational Measurement Center

An extensive literature exists concerning construction, validation, and use of mental tests. In view of this fact, there is a surprising lack of research concerning the actual building blocks of objective mental tests--test items. Wesman (1971) has pointed out that texts on measurement are filled with rules of thumb for construction of test items, yet little empirical research exists which validate these rules. Only sparse information is available concerning effects of violating these rules for item-writing or the strictness with which these rules should be followed.

One aspect of test item construction which has received some attention in the literature is item clueing. A multiple-choice item is clued when structural or grammatical faults aid unknowledgeable examinees to find the correct answer. Several studies have examined grammatical mismatches between incorrect answers and the stem and items in which the correct answer is much longer or shorter than the distractors (Board & Whitney, 1972, and McMorris, Brown, Snyder, & Purzell, 1972, for example). The results of this research have been inconsistent.

Rules concerning test item formats, such as those suggested by Adkins (1974), have also received some study. Several investigators have studied items which contain responses such as "none of above" or "all of above" (Rimland, 1960; Hughs & Trimble, 1965; Williamson & Hopkins, 1967). Items containing alternatives such as these tend to be more difficult, but do not differ from control items in other respects. Furthermore, these results have been inconsistent. Board and Whitney (1972) and Vitola and Cantrell (1961) have studied open-stem items--items in which responses complete the stem, as contrasted with items in which the stem is a grammatically complete question. In both of these studies, open-stem items were found to be more difficult than closed-stem items. Cieutat (1960) found that rote-memory items had both higher reliabilities and validities than application items.

As can be seen, the research results concerning format rules for multiple-choice test items have been both sparse and inconsistent. The purpose of the present study is to provide further information concerning the usefulness of several rules that have been proposed for test item formats. Four item-writing rules are examined here. These item formats to be tested are (a) multifactor items--items in which the correct answer contains several parts, (b) situational items--items which require application of knowledge to real-life situations, (c) negative items--items in

which the examinee must select the one incorrect answer, rather than the one correct answer, and (d) open-stem items. Dependent variables examined are item difficulty and reliability. For each item which was constructed to have one of the experimental formats in this study, a control item was written covering parallel content but which was single-factor, nonsituational, positive, and closed-stem. Therefore, item content was controlled in tests of the experimental formats. This design allowed a purer test of the experimental formats.

Method

Twenty-five four-alternative multiple-choice items were constructed using each of the four experimental formats. The content of these items was general information in areas such as science, math, English, art, and history. For each experimental item, a control item was written covering, as much as possible, the same content. The control items followed none of the formats to be tested. Instead, they were positive, single-factor, closed-stem, and nonsituational. Tables 1 through 4 contain examples of experimental and matching control items for each format tested. Two 100-item test forms were compiled from these experimental and control items. Each item appeared on only one form. The experimental items for each format were divided evenly between the two forms. An experimental item and its matching control item always were placed on opposite forms, but in the same sequence location.

Each test form was administered to approximately 100 airmen who held grades between E-4 and E-6. The tests were administered under anonymous conditions. Therefore, no data is available concerning the age, sex, or ethnic group characteristics of the samples.

Results

For each item, the difficulty was computed by determining the proportion of airmen who selected the correct answer. Three reliability estimates were computed for each item. First, the point-biserial correlation was computed between the score on each item and the total score for the test form on which the item appeared. Then each test form was factor-analyzed. On each test form, an inspection of the eigenvalues revealed discontinuities at 6 and 11 factors. Therefore, varimax rotation was applied to both 6- and 11-factor solutions for each test form. Squared multiple correlations were used as communality estimates. For each item, the factor was identified in each rotated solution with which the item had the largest correlation. These largest correlations were used as additional estimates of the items' reliabilities.

Therefore, one difficulty three reliability estimates were available for each item. For each experimental item format, an analysis of variance (ANOVA) was done with each of these four dependent variables. The reliability estimates were squared before being analyzed. The unit of analysis in these ANOVAs was the individual test item. The independent

variables were experimental vs control and test form. Because experimental and control items were matched on content but always appeared on different forms, a Latin square design was used in the ANOVAs. The layout of this design is illustrated in Figure 1. This design permits both a between-groups test and within-groups tests of main effects for the two independent variables. Interaction effects could not be tested with this design. Table 5 contains percents of variance attributable to each effect for the various ANOVAs and degrees of freedom. Out of 48 F-ratios computed in the ANOVAs, only one was significant at the .05 level. This is probably a chance result.

Discussion and Conclusions

The present results indicate that none of the four test item formats examined influence either difficulty or reliability, after item content is controlled. Findings of Vitola and Cantrell (1961) concerning open-stem items and of Cieutat (1960) concerning application items were not replicated.

The present results show that the four item formats tested should be neither preferred nor avoided in order to control difficulty or reliability. However, the present results do not preclude the possibility of format-by-content interaction effects on these dependent variables. Some formats may be more appropriate for some content. Furthermore, other considerations may dictate preference for or avoidance of these formats in test construction. For example, situational items may have greater content validity than rote memory items and therefore might be preferred when content validity is important.

References

Adkins, D. Test Construction. Columbus, Ohio: Merrill, 1974.

Board, C., & Whitney, D. R. The effect of selected poor item-writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 1972, 9, 225-233.

Cieutat, V. J. Reliability and validity of two types of objective test items. Psychological Reports, 1960, 7, 447-449.

Hughs, H. H. & Trimble, W. E. The use of complex alternatives in multiple-choice items. Educational and Psychological Measurement, 1965, 25, 117-126.

McMorris, R. F., Brown, J. A., Snyder, G. W., & Purzell, R. M. Effects of violating item construction principles. Journal of Educational Measurement, 1972, 9, 287-295.

Rimland, B. The effects of varying time limits and of using "right answer not given" in experimental forms of the U. S. Navy arithmetic test. Educational and Psychological Measurement, 1960, 20, 787-794.

Vitola, B. M. & Cantrell, G. K. An Experimental Investigation of Multiple-Choice Item Structure. PRL-TDR-61-49. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, 1961.

Wesman, A. G. Item writing. In Thorndike, R. L. (ed.), Educational Measurement, 2nd ed., American Council on Education, 1971.

Williamson, M. L. & Hopkins, K. D. The use of "none of these" versus homogeneous alternatives on multiple-choice tests: experimental reliability and validity comparisons. Journal of Educational Measurement, 1967, 4, 53-58.

Table 1

Sample Test Items - Multifactor

Experimental	Control
Ordinary table salt is made up of what elements? a. Chlorine and magnesium b. Potassium and magnesium c. Potassium and sodium d. Chlorine and sodium	Ordinary table salt contains which one of the following elements? a. Chlorine b. Fluorine c. Magnesium d. Potassium
If $2x/y = 10$, which of the following values for x and y are correct? a. $x = 10$, $y = 3$ b. $x = 14$, $y = 7$ c. $x = 15$, $y = 5$ d. $x = 20$, $y = 4$	If $2x/4 = 10$, what value for x is correct? a. 10 b. 14 c. 15 d. 20
What state was the most recent to enter the Union, and in what year did it enter? a. Alaska in 1959 b. Hawaii in 1959 c. Alaska in 1960 d. Hawaii in 1960	In what year was the most recent state admitted to the Union? a. 1958 b. 1959 c. 1960 d. 1961

Table 2

Sample Test Items - Negative

Experimental

Which one of the following animals is NOT a vertebrate?

a. Humming bird
b. Rattlesnake
c. Tuna fish
d. Earthworm

Which one of the following plays was NOT written by William Shakespeare?

a. Pygmalion
b. The Tempest
c. Julius Caesar
d. The Merchant of Venice

Which one of the following works of art was NOT produced by Michelangelo?

a. David
b. La Pieta
c. The Last Supper
d. The Sistine Ceiling

Control

Which one of the following animals is a vertebrate?

a. Octopus
b. Starfish
c. Earthworm
d. Rattlesnake

Which one of the following plays was written by William Shakespeare?

a. Faustus
b. Pygmalion
c. The Tempest
d. The Crucible

Which one of the following works of art was produced by Michel-angelo?

a. The Mona Lisa
b. Venus De Milo
c. The Last Supper
d. The Sistine Ceiling

Table 3

Sample Test Items - Open Stem

Experimental	Control
The largest planet in our solar system is	What planet is the largest in our solar system?
a. Jupiter	a. Jupiter
b. Neptune	b. Neptune
c. Uranus	c. Uranus
d. Saturn	d. Saturn
Suffrage is the right to	Suffrage refers to what right?
a. petition	a. To petition
b. assemble	b. To assemble
c. worship	c. To worship
d. vote	d. To vote
The predominant religion in Chile is	What religion is predominant in Chile?
a. Episcopalian	a. Episcopalian
b. Methodist	b. Methodist
c. Lutheran	c. Lutheran
d. Catholic	d. Catholic

Table 4

Sample Test Items - Situational

Experimental	Control
If a 15-ounce bottle of dish-washing liquid cost $.90, what is the cost per ounce of the liquid? a. 4 cents b. 6 cents c. 10 cents d. 15 cents	What is 1/15 of .9? a. .04 b. .06 c. .10 d. .15
A student wishes to find information in a library on the current state of the economy in the U. S. Which one of the following sources would be most useful in finding this information? a. An encyclopedia b. The Readers' Guide c. The card catalogue d. An economics textbook	What material is contained in the Readers' Guide? a. Condensed novels b. Famous quotations c. Definitions of words d. Titles of magazine articles
If a person planned a trip to Brazil, what language course would probably be most useful to him or her? a. English b. Spanish c. French d. Portuguese	What is the official language of Brazil? a. Spanish b. English c. French d. Portuguese

Table 5

ANOVA Results - Proportions of Variance Attributable to Various Sources

Situational	diff	R1	R6	R11	df
Between item contents	.8403	.8379	.5856	.4806	25
item groups	.0617	.0927	.0025	.0048	1
items within groups	.7786	.7452	.5831	.4758	24
Within item contents	.0097	.0003	.0169	.0566	50
test form	.0053	.0004	.0169	.0389	1
format (exp vs con)	.0044	.0000	.0001	.0177	1
Residual	.1500	.1618	.3975	.4628	48

Open Stem	diff	R1	R6	R11	df
Between item contents	.9551	.7285	.6172	.6752	25
item groups	.0939	.0699	.0434	.0619	1
items within groups	.8612	.6586	.5738	.6133	24
Within item contents	.0003	.0069	.0642	.0190	50
test form	.0001	.0049	.0642	.0187	1
format (exp vs con)	.0002	.0020	.0000	.0003	1
Residual	.0446	.2646	.3186	.3058	48

Table 5 (continued)

ANOVA Results - Proportions of
Variance Attributable to Various Sources

Negative	diff	R1	R6	R11	df
Between item contents	.7067	.7191	.5578	.6440	25
item groups	.0287	.0008	.0244	.0470	1
items within groups	.6780	.7183	.5334	.5970	24
Within item contents	.0685	.0133	.0362	.0347	50
test form	.0573	.0039	.0250	.0264	1
format (exp vs con)	.0133	.0089	.0126	.0096	1
Residual	.2248	.2676	.4060	.3213	48

Multifactor	diff	R1	R6	R11	df
Between item contents	.8591	.8671	.6100	.4963	25
item groups	.0007	.0098	.0019	.0257	1
items within groups	.8584	.8573	.6081	.4706	24
Within item contents	.0110	.0093	.0005	.0169	50
test form	.0000	.0000	.0002	.0163	1
format (exp vs con)	.0110*	.0093	.0003	.0007	1
Residual	.1299	.1236	.3900	.4868	48

*$p < .05$

Figure 1

Analysis of Variance Design

		Experimental	Control
item group 1	content 1	form A	form B
	content 2	form A	form B
	.	.	.
	.	.	.
	.	.	.
	content 13	form A	form B
item group 2	content 14	form B	form A
	content 15	form B	form A
	.	.	.
	.	.	.
	.	.	.
	content 25	form B	form A

Acknowledgements

The authors wish to thank 2Lt Scott Gregoire, who played a key role in the construction of the experimental tests, and A1c Linda A. Jones, who provided able assistance in the gathering and analysis of the data reported here. In addition, much appreciation is due to the personnel at the USAF Occupational Measurement Center and at other organizations of Lackland AFB, without whose cooperation and assistance this research could not have been completed.

THE POLYCHOTOMOUS RESPONSE MEASUREMENT MODEL AND THE FREQUENCY RATIO METHOD

by

William J. Moonan, Ph.D
Consultant for Statistics

Navy Personnel Research and Development Center
San Diego, California 92152

A Presentation Made to
The 1977 Military Testing Association Conference
San Antonio, Texas

17-21 October 1977

I. INTRODUCTION

The purpose of this report is to define the components and analysis of a mental measurement model which may be used to analyze responses classified in one of several categories of the items of the measurement instrument. Such measurement instruments may be the usual tests or examinations (with extended response alternatives) or questionnaires. This situation also covers performance measurement when it is desired to have extended categories of correct or incorrect response.

To fulfill the purpose we first present some background material to the approach we use to specify the model and its parameters. The method used to estimate the parameters is developed as a generalization of a psychometric estimation procedure, called the Frequency Ratio Method, devised by the author for analyzing the Binary Response Measurement Model. Applications are suggested and an example is provided.

II. BACKGROUND

2.1 Influences of G. Rasch, B. D. Wright and P. Blommers

After the publication of a new book by Rasch (1960) on measurement models, this work received little attention at measurement centers and universities of the U. S. except at the State University of Iowa and the University of Chicago.

Since Rasch's work was unique, interesting, provocative, and useful, considerable speculation as to why the book did not generate more attention is justified. The most spectacular aspect of Rasch's models were their "specifically objective" features meaning, for example, that a person's ability, as estimated from his responses, is independent of the items used to measure the ability, and the easiness or difficulty of the items is independent of the populations from which the persons are sampled. Hence, Wright succinctly refers to Rasch Models with "Person free item measurement and item free person measurement."

The truth of this property has often been challenged, but most challengers forget that these properties are easily proved mathematically and are empirically true,

a. only if the data fit the measurement model being tested, and

b. only within measurement error of the statistics involved.

My thought on this issue is that many of the readers of Rasch's work suffered from "Objectivity Shock" meaning that they did not consider Specific Objectivity possible and put the book aside. Neither Professor Ben Wright of the University of Chicago nor Professor Paul Blommers of the University of Iowa were greatly bothered by this problem but did their own research in this type of measurement and encouraged Ph.D. candidates to do likewise. This leadership greatly popularized Rasch's work in this country and served to develop it further. Wright also taught courses on Rasch Models at professional meetings. Most of Wright's work as well as that of his students dealt with Rasch's achievement test model whereas Blommer's students worked with that model as well as Rasch's models for reading speed and oral reading accuracy.

Wright and Panchapakesan (1969) developed an unconditional maximum likelihood estimation procedure and computerized the analysis for the Rasch achievement model.

2.2 The Binary and Polychotomous Measurement Models

In most achievement and ability tests and examinations, the responses to the test items are classified into one of two categories--either "correct" or "incorrect." This circumstance leads to a metrological model for analysis which may be termed a "Binary Measurement Model" or BMM. The BMM is also called a dichotomous model. It is this model that Wright (1968), Rasch (1960) and Moonan (1974) considered and provided different, but equivalent, analysis for.

However, there are many other response situations that are useful and must be accommodated and analyzed. One of these occurs in achievement and ability testing and gives rise to responses, classified by testees or psychometricians, into one of several (r>2) categories. This situation naturally occurs in performance measurement and in attitude, interest, and motivation surveys and in questionnaires. In such cases the Binary Response Measurement Models (BMM) are not appropriate and a generalization is required. It is our purpose to define an analysis for such a model, called the Polychotomous Response Measurement Model (PMM), in this report. Such models have been considered by others: Rasch (1961), Vogt (1971) and Anderson (1973). Anderson has provided a successful computerized analysis of the PMM. This used a new and very complicated mathematical estimation procedure called "Conditional Maximum Likelihood." The approach used here is greatly simplified and is an extension used by Moonan (1974) to analyze the BMM. Vogt made her analysis with unconditional maximum likelihood techniques, but indicated her approach was not entirely successful (personal communication). Rasch did not carry the analysis of the PMM very far.

2.3 The Process of Mental Metrology

We shall attempt now to outline the considerations believed important for developing and analyzing a mental metrological instrument.

2.3.1 The Instrument Goal and Objectives

The first thing for the metrologist to do is to decide what mental quality or characteristic of persons he desires to measure. This is called his "metrological goal" and indicates the general property that the instrument is designed to measure. Other related sub-characteristics intended to be also measured by the same device are called "objectives."

2.3.2 The Items and Their Functions

The measurement instrument may be thought of as a kind of large "hurdle," the success in surmounting of which by a person, measures his skill at the task attempted. Rather than having one single hurdle it is customary for the instrument to have several, I, related small hurdles, called items, each of which is designed to

(a) elicit a piece of relevant information, knowledge or skill, or

(b) provide an indication of an attitude, interest, opinion or judgment regarding a problem, situation or circumstance related to the goal and objectives.

Each item is assumed to have a property, symbolized by $\epsilon(i)$, and called its easiness (or difficulty, $\delta(i)$), which characterizes the facility with which the most favorable responses to these items are evoked from persons in general. Occasionally an item parameter, called the index of discrimination, is used in the model. Both Schmidt (1969) and Wright (1968) have indicated that the use of this index in the model is fatal to the achievement of objectivity, so we omit its consideration in our system.

2.3.3 Responses - Free and Controlled

Each item has associated with it, a single or set of $\underline{R \geq 2}$ replies, called responses, appropriate to the problem or question posed by the item. It is most efficient if these responses are indicated by the same categories for each item. Achievement or ability measuring instruments usually have Binary Categories such as

(a) Correct and Incorrect; R=2

(b) True and False ; R=2

(c) Yes and No ; R=2

(d) Right and Wrong ; R=2

Instruments measuring affect usually have either Polychotomous or Binary categories such as

(a) like, indifferent, dislike ; R=3

(b) strongly agree, agree, undecided, disagree, strongly disagree ; R=5

(c) approve, disapprove ; R=2

(d) True, False ; R=2

(e) yes, maybe, no ; R=3

(f) synonym, antonym, neither ; R=3

(g) satisfactory, unsatisfactory ; R=2.

Responses to some item types are called "Free" if this reply is uncontrolled or not constrained in any way but may assume any form, character or specificity including written or verbal phrases, or statements, or numerical calculations. On the contrary, controlled responses imply a set of categories, the same or not for each item, into one of which the natural reply of the person to the item may be classified or designated by the person or the psychometrician.

The response categories are often ordered and assumed orderable on a scale of correctness, appropriateness, suitability, desireability or some other quality appropriate to the goals and objective. This scale is calibrated and a point for each response category is indicated by $\theta(r)$, $r=1,R$. We assume that the least desireable category is $r=R$ for which $\theta(R)=0$.[1] This assumption is made for both Binary and Polychotomous Response models. Because of the importance of this assumption in our subsequent derivations we designate it separately from the text as

(1) $\theta(1) = 1$
$\theta(R) = 0.$

In the dichotomous model the $\theta(r)$ are considered to be the observables $X(n,i) = 0$ or 1. Although this could also be done for the polychotomous model where $X(n,i,r)$ represents the response category related, we do not choose to do so because we wish to identify the response category parameters to be quantified via analysis.

Early analysis in psychometrics routinely made the assumption given by (1). Since metrology models were not employed their only reason for assuming (1) was "convenience". There is, of course, a more profound and necessary reason which we shall discover.

2.3.4 Person Parameters

Each person receives a "score" (to be defined later) on the instrument from which a value, $\alpha(n)$, of the characteristic being measured, may be estimated. The estimation of $\alpha(n)$ for person n is usually the single purpose for the psychometrician to administer the instrument to the person. The $\alpha(n)$ for some scores are non-estimable. Were a person to receive the "best or worst" score that it is possible to achieve on the instrument it would not be appropriate to estimate $\alpha(n)$ in either case because the instrument is either two easy or difficult for such persons and the appropriate $\alpha(n)$ is too uncertain. Consequently we assume that

(2) $\alpha(n)$ for "perfect" or "zero" scores is non-estimable.

2.3.5 Psychometric Models

We have noted the desireability of considering the parameters,

(a) Item parameters, $c(i)$; $= 1,I$

(b) Response parameters, $\theta(r)$;$r = 1,R$

(c) Person parameters, $\alpha(n)$;$n = 1,N$

A measurement model establishes a relation or relations among the responses, $X(n,i,r)$, and the parameters of the model. If $r=1,2=R$ the model is called a Binary Measurement Model, BMM. If $R>2$ the model is called a Polychotomous Measurement Model, PMM. Measurement models are of varying degrees of complexity and form, see Lord and Novick (1968), depending upon the relations assumed and the number and nature of the parameters involved.

2.3.6 Model Analysis

The measurement models form the basis for finding estimates of parameters desired to be estimated and for making a statistical test of agreement between the model and the responses. Model analysis means estimating the parameters and testing the agreement. The techniques for making analysis of measurement models includes various statistical and mathematical procedures such as

(a) least squares

(b) conditional and unconditional maximum likelihood

(c) simple statistical analysis

(d) analysis of variance

(e) probability analysis, and

(f) algebraic analysis.

III. THE POLYCHOTOMOUS MEASUREMENT MODEL (PMM)

3.1 Odds and Ends

The odds the person n will respond to item i with response r is defined to be

(3) $$O(n,i,r) = [a(n)c(i)].^{\theta(r)} \qquad \begin{array}{l} n = 1,N \\ i = 1,I \\ r = 1,R \end{array}$$

The sum of the odds for that person to make a response other than r on item i is

(4) $$O(n,i,\bar{r}) = \sum_{k}^{R} O(n,i,k) \qquad k \neq r.$$

The idea of developing measurement models in probabilistic terms via expressions for the odds of a specific response is due to Rasch (1960).

From elementary probability theory we know that if the odds for the occurrence of an event E are a/b, then the probability of the event E is

(5) $$P(E) = \frac{a}{a+b}.$$

Now let E be the event characterized in our notation by (n,i,r) and consider the odds ratio

(6) $$\frac{a}{b} = \frac{O(n,i,r)}{O(n,i,\bar{r})}$$

As the ratio of odds that person n will make response r to item i to the odds that he will make some other response to that item. The probability of this event is, by (5), equal to

$$(7) \qquad P(n,i,r) = \frac{O(n,i,r)}{\sum_{r}^{R} O(n,i,r)} = \frac{O(n,i,r)}{O(n,i,.)} .$$

This is a form often assumed by other workers for the PMM. We could have started here to satisfy those purists who do not believe we are justified in applying (5) to (6) in order to get (7). It is more convenient to work with probabilities than odds because the calculus of odds is not as well developed as the Theory of Probabilities.

Note that our model is unidimensional in the sense that we do not have a separate dimension of ability and easiness for each response category as was postulated by Anderson (1973), and Vogt (1971). Such a multivariate model may be useful, upon occasion, but is difficult to interpret and use in practice so we do not employ it. Our problem is difficult enough without such complications and we are content ot start with this assumption, rather than, perhaps, make it in the end anyway. The multivariate parameters may be expressed as

$$(8) \qquad [\alpha(n)]^{\theta(k)} = \alpha\binom{k}{n} \text{ and } [\varepsilon(i)]^{\theta(k)} = \varepsilon\binom{k}{i}$$

which, in our model and notation, are expressed as

$$(9) \qquad \ln \alpha\binom{k}{n} = t\binom{k}{n} = \theta(k)t(n) \text{ and } \ln \varepsilon\binom{k}{i} = \theta(k)t(i) = t\binom{k}{i}$$

The problem with our simplification is that it may be too simple a model to adequately represent some sets of response data. In that case the researcher should consult Anderson (1973) or use another approach such as that provided by Bock (1970).

3.2 Applying the Frequency Ratio Method to the PMM

In 1974 the author developed an analysis for a simpler model, the BMM. This procedure is called the Frequency Ratio Method because certain probabilities, analogous to (7), were estimated as objective probabilities using observed frequencies. We employ the same approach here since we believe it to be easier, and almost as accurate, as the estimation procedures mentioned on page 5.

Consider now the four possible events which can occur if the same person n responds to items i and j with either response s or response r. We have, since the responses are assumed independent, see (7),

$$(10) \qquad P[(n,i,s),(n,j,s)] = \frac{O(n,i,s) \cdot O(n,j,s)}{O(n,i,\cdot) \cdot O(n,j,\cdot)} ,$$

$$(11)\qquad P[(n,i,s),(n,j,r)] = \frac{O(n,i,s)\cdot O(n,j,r)}{O(n,i,\cdot)\cdot O(n,j,\cdot)}\ ,$$

$$(12)\qquad P[(n,i,r),(n,j,s)] = \frac{O(n,i,r)\cdot O(n,j,s)}{O(n,i,\cdot)\cdot O(n,j,\cdot)}\ \text{, and}$$

$$(13)\qquad P[(n,i,r),(n,j,r)] = \frac{O(n,i,r)\cdot O(n,j,r)}{O(n,i,\cdot)\cdot O(n,j,\cdot)}\ .$$

Using (3) we can express (11)-(12) in parametric terms with (14), as (15) and (16).

$$(14)\qquad D(i,j,\cdot) = O(n,i,\cdot)\cdot O(n,j,\cdot)$$

$$(15)\qquad P[(n,i,s),(n,j,r)] = [\alpha(n)\varepsilon(i)]^{\theta(s)}[\alpha(n)\varepsilon(j)]^{\theta(r)}\Big/D(i,j,\cdot)$$

$$(16)\qquad P[(n,i,r),(n,j,s)] = [\alpha(n)\varepsilon(i)]^{\theta(r)}[\alpha(n)\varepsilon(j)]^{\theta(s)}\Big/D(i,j,\cdot).$$

It is interesting and comforting to note that for the BMM with $\theta(s=1) = 1$ and $\theta(r=4) = 0$ expressions (15 and (16) agree identically with equations (7) and (8) in Moonan (1974).

The conditional probability that person n responds to item i with response s given that he used both responses s and r for the two items i and j is

$$(17)\qquad P[(n,i,s)\,|\,\{((n,i,s)\cap(n,j,r))\cup((n,i,r)\cap(n,j,s))\}] = \frac{(15)}{(15)+(16)}\ ;$$

also

$$(18)\qquad P[(n,i,r)\,|\,\{((n,i,s)\cap(n,j,r))\cup((n,i,r)\cap(n,j,s))\}] = \frac{(16)}{(15)+(16)}\ .$$

Simplifying (17) and (18) using (3) we note that $\alpha(n)$ cancels showing that the PMM has the same and usual property of objectivity that the BMM and other Rasch models enjoy. This means that for response data sets that fit (7) the estimated parameters of $\varepsilon(i)$ and $\theta(r)$ are invariant estimates, except for measurement error, independently of the populations from which the persons are obtained, and from which the items were selected.

We can rewrite (15) and (17) using (3) as

$$(19)\qquad P[{}^{ij}_{sr}] = \frac{[\varepsilon(i)]^{\theta(s)}\cdot[\varepsilon(j)]^{\theta(r)}}{[\varepsilon(i)]^{\theta(s)}\cdot[\varepsilon(j)]^{\theta(r)}+[\varepsilon(i)]^{\theta(r)}\cdot[\varepsilon(j)]^{\theta(s)}},\qquad j>i$$

and (16) and (18) as

$$(20) \qquad \begin{bmatrix} P^{ij}_{rs} \end{bmatrix} = \frac{[\epsilon(i)]^{\theta(r)} \cdot [\epsilon(j)]^{\theta(s)}}{[\epsilon(i)]^{\theta(s)} \cdot [\epsilon(j)]^{\theta(r)} + [\epsilon(i)]^{\theta(r)} \cdot [\epsilon(j)]^{\theta(s)}} , \qquad i>j$$

Note that (19) and (20) are independent of $\alpha(n)$ and their denominators are identical. Using response data we can estimate the probabilities represented by (19) and (20) with observables. If we take the ratio of (19) and (20) we get the considerable simplification

$$(21) \qquad \frac{P\begin{bmatrix} ij \\ sr \end{bmatrix}}{P\begin{bmatrix} ij \\ rs \end{bmatrix}} = \frac{[\epsilon(i)]^{\theta(s)-\theta(r)}}{[\epsilon(j)]^{\theta(s)-\theta(r)}}$$

3.3 From Probability to Frequency

Let,

(22) $f(i,s;j,r)$ be the observed frequency with which persons responded to item i with response s, and to item j with response r. Similarly, we define $f(i,r;j,s)$ and $g(i,j;r,s) = f(i,s;j,r) + f(i,r;j,s)$

Now (19) may be empirically estimated by

(23) $f(i,s;j,r)\Big/g(i,j;r,s)$.

Similarly (20) is estimated by

(24) $f(i,r;j,s)\Big/g(i,j;r,s)$,

and hence (21) is simply estimated by the frequency ratio

(25) $f(i,s;j,r)\Big/f(i,r;j,s)$

3.4 Item and Response Parameter Estimation

Given the response data the frequencies for all pairs of items and responses may be easily tabulated. From these data the frequency ratios

(26) $R(i,j) = \dfrac{f(i,s;j,r)}{f(i,r;j,s)}$ for $i>j = 1,I$

(27) $R(j,i) = \dfrac{f(i,r;j,s)}{f(i,s;j,r)}$ for $j>i = 1,I; s\neq r = 1,R$

and,

(28) $R(i,i) =$ Blank.

may be easily computed. We note that (26) estimates

(29) $$\frac{[c(i)]^{\theta(s)-\theta(r)}}{[c(j)]^{\theta(s)-\theta(r)}}$$, and (27) estimates

(30) $$\frac{[c(j)]^{\theta(r)-\theta(s)}}{[c(j)]^{\theta(r)-\theta(s)}}$$. Also let

(31) $$\theta(s)-\theta(r) = \Delta(s,r).$$

We now may express the natural logarithm of (29) and (30) as

(32) $$\Delta(s,r)\ln c(i) - \Delta(s,r)\ln c(j) = \ln R(i,j), j>i, \text{ and}$$

(33) $$-\Delta(s,r)\ln c(j) + \Delta(s,r)\ln c(i) = -\ln R(i,j), i>j.$$

Now if we add (32) and (33) over i for fixed j, after adding and subtracting $\Delta(s,r)\ln c(j)$ we get

(34) $$-\Sigma\ln R(i,j) = -T(i) = I\Delta(s,r)\ln c(i) - \Delta(s,r)\sum_{j}^{I}\ln c(j)$$

We now make the final assumption for our method. For definiteness of the c(i) scale, we assume that

(35) $$\sum_{j}^{I}\ln c(j) = \sum_{i}^{I}\ln c(i) = 0.$$

From (34) we may now write

(37) $$t(i) = \frac{-T(i)}{I} = \Delta(s,r)\ln c(i).$$

Because $\Delta(s,r)$ is a function of $\theta(r)$ and $\theta(s)$ we have difficulty in estimating c(i) explicitly from (37). Fortunately we foresaw this difficulty and made assumption (1). Setting s=1 and r=R, then $\theta(s) = 1$, $\theta(r) = 0$ and $\Delta(s,r) = \theta(s) - \theta(r) = 1-0 = 1$. We now write (37) as

(38) $$\ln c(i) = t(i)$$, or

(39) $$c(i) = e^{t(i)}$$.

Let the c(i) as estimated from response categories (1,k), k≠R be designated as $c\binom{k}{i}$, then by (9) we write

(40) $$t\binom{k}{i} = \theta(k)t(i). \quad k\neq 1,R$$

Equation (40) enables us to estimate the response parameters $\theta(k)$ having first estimated the $\varepsilon(i)$ using data for response categories 1 and R. Since this process may result in considerable estimation error we will use an averaging process in practice. This will be demonstrated in our examples.

3.5 Person Parameter Estimation

Our analysis has shown the theoretical derivation of the item parameters $\varepsilon(i)$ and the response parameters $\theta(r)$. We need next to consider the estimation of $\alpha(n)$ for a person. To do this we introduce the concept of a "score".

Let the vector of estimated response parameters be denoted by θ and its transpose by θ', then

(41) $$\theta = (1,\theta(2),\theta(3), \ldots , 0)$$

and let Y be the "score vector" of integers $y(k)$ indicating the number of the I items responded to in the kth response category by the nth person. Then

(42) $$Y = (y(1),y(2), \ldots , y(k)), \text{ where}$$

(43) $$\sum_{k}^{R} y(k) = I$$

If person n does not respond to all I items because of time or other reasons, (43) will not be strictly true. This does not bother our theory, but merely extends the computations.

We now define the "score" of a person to be

(44) $$S(n) = Y\theta'.$$

This definition is congruent with the usual definition of scores in the Binary Model where the score is the number of correct responses. For example, suppose I = 20 and person n got 12 items correct and 8 items incorrect. Then

(45) $$S(n) = (12,8)\binom{1}{0} = 12$$

In binary scored tests the number of possible scores is I+1, but for polychotomous tests the number of possible scores, even with (43), can be very large. This number, S, is the total number of permutations of each part of the possible partitions of I, among R categories.

For example suppose I=5 and R=4, then the partitions of 5 that are of interest when R=4 are given below for the parts of (46). In this case

$$S = 4+12+12+4+12+12 = 56$$

5000*	0104	2111	0023
0500	0014	1211	0203
0050	3110	1121	2210
0005*	3101	1112	2201
4100	1130	3200	2012
4010	1031	3020	0221
4001	0131	3002	0212
1400	1310	2300	0122
0410	1301	0320	1202
0401	0311	0302	1022
1040	1103	2030	2102
0140	1013	0230	1220
0041	0113	0032	2120
1004	3011	2003	2021

* non-estimable

Figure 1. Possible score vectors for the polychotomous situation with l=5, k=4.

$$
(46) \quad \begin{array}{ll} 5,0,0,0 & 2,1,1,1 \\ 4,1,0,0 & 3,2,0,0 \\ 3,1,1,0 & 2,2,1,0 \end{array}
$$

Permuting each part of each partition among R=4 categories we get the list of S=56 possible score vectors shown in Figure 1. We assume here that (43) holds otherwise the list is considerably larger.

In order to estimate the $\alpha(n)$ associated with a given score we find the $\alpha(n)$ for which the sum of (7), for response 1, over the items which equals the score, i.e.

$$
(47) \qquad \sum_{i}^{I} P(n,i,1) = S(n).
$$

Equation (47) is solved for $\alpha(n)$, recalling (7) and (3), by iteration such as by the Method of Newton-Raphson. Accordingly, for each score vector of interest the associated $\alpha(n)$ may be determined. Notice also that in this procedure no infromation regarding which items were responded to in the manner indicated by the score vector is utilized.

In order to estimate the ability, $\alpha(n)$, associated with the score $S(n)$ for our model we write

$$
(48) \qquad \sum_{i} \frac{[\alpha(n)\varepsilon(i)]}{\sum_{r}[\alpha(n)\varepsilon(i)]^{\theta(r)}} = S(n)
$$

as the function upon which the estimation of $\alpha(n)$ is to be based. Our principle of estimation is that $\alpha(n)$ is to be derived by setting $S(n)$ equal to its expectation in the model. This is accomplished by using (48).

Our problem is to find that $\alpha(n)$ that maximizes (48). This problem is efficiently solved by using the Newton-Raphson method of finding a root of an equation of the form $f(x) = 0$. To this end we rewrite (48) as

$$
(49) \qquad f(x) = \sum_{i} \left[\frac{x\varepsilon(i)}{\sum[x\varepsilon(i)]^{\theta(r)}} \right] - S(n) = 0
$$

We start by guessing at the root, the guess being $x = \alpha(n) = x_1 = S(n)/[I-S(n)]$, say. Suppose, however, that the root is actually x_1+h. Then by Taylor's calculus expansion of $f(x)$,

$$
(50) \qquad f(x_1+h) = f(x_1) + hf'(x_1) + \dots
$$

Our reasoning takes the following line:

1. The error of approximation to the real root is h.

2. The error is small enough to permit us to ignore the terms in the Taylor expansion in which "h" appears to a power greater than the first.

3. We may therefore write, since the actual root is (x_1+h),

(51) $0 = f(x_1) + hf'(x_1)$ approximately.

The error is,

(52) $h = -f(x_1)/f'(x_1)$ approximately.

The iterated use of this method gives the general formula for the approximation to x as

(53) $$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

This iterative procedure is easily programmed for digital computers once the derivative of f(x) has been made explicit. Moonan and Potterton have done this at NPRDC. This subroutine can be used to develop a table or individual values of $\alpha(n)$ associated with scores S(n) given vectors of values for the $\varepsilon(i)$ and $\Theta(r)$ for a particular measuring instrument.

3.6 Agreement Between the Model and the Response Data

We can now assume that we have response data for which the item parameters $\varepsilon(i)$ and the response parameters $\Theta(r)$ have been estimated. We desire next to use these estimates to test if the model, represented by (7), agrees with the observed response data. This test is sometimes referred to as a "Goodness of Fit Test". The hypothesis tested is that the items used evoked responses from the persons in accordance with the PMM. To make this test we use Fisher's (1948) χ^2 goodness of fit test. To this end consider

(54) $F(\begin{smallmatrix} i \\ j \end{smallmatrix})$ = A matrix of the observed frequencies f(i,s;j,r) and f(i,r;j,s), defined by (22), and arranged in I rows and I columns for fixed s and r.

For every pair of i and j both (19) and (20) can be computed with the parametric estimates $\varepsilon(i)$ and $\Theta(r)$. When these are appropriately multiplied by g(i,j,r,s) of (22) they provide expected frequencies, under the model, with which $\chi^2(i,j)$ may be computed and tabulated to get a $\chi^2(i)$ for each item with I-R degrees of freedom. This test also serves as a basis for selecting items to include or not in the analysis. See Moonan (1974, p. 13).

IV. APPLICATIONS

4.1 Applications to Ability Testing

Measurement of Abilities with the BMM is common practice and need not be elaborated upon here except to say that in this type of testing binary response categories are often too restrictive. It would be better, for some purposes, if the response categories could be extended. For example in performance testing the method used to find an answer may be as important as the answer itself. In such a case four response categories may be appropriate which we diagram as follows:

	correct answer	incorrect answer
correct method	r = 1	r = 2 or 3
incorrect method	r = 2 or 3	r = 4

Whether the upper right cell or the lower left cell corresponds to r = 2 or r = 3 may be data dependent.

4.2 Applications to Attitude Estimation

An attitude instrument is often designed to appraise a person's favorableness toward some institution, group or concept. Other types measure interests and motivations. Often the response categories are polychotomous and consist of prescaled statements indicating degrees of significance or a status condition. These devices seem most naturally to be analyzed with a PMM.

4.3 Applications to Surveys and Questionnaires

If all or part of survey or questionnaire instruments seeks to measure a unidimensional property of persons <u>and</u> the responses can be transformed or converted to the same set for each question, then the PMM is applicable and the analysis given here is appropriate. An example of this is a survey instrument designed to assess the quality of houses for the purpose of establishing a value for sale or for insurance. It seems better to establish the uniform response categories a priori rather than go through a re-coding process.

V. EXAMPLE

5.1 A Simulation Example

We have not yet completed a computer program that carries out the entire PMM analysis. We expect to do this in the future. In the meantime we show our procedure by means of a very detailed example using simulated data. We organize the computations in five phases, as follows:

	Phase	
	I	Data Tabulations
(55)	II	$\varepsilon(i)$ Estimation
	III	$\theta(r)$ Estimation
	IV	Goodness of Fit
	I	$\alpha(n)$ Estimation

Each phase involves several steps.

5.1.1 The Situation

Although real data for the PMM are plentiful we chose to simulate the set we use here because, in doing so, we would then know the true values of the parameters we try to estimate. The advantages of this practice are obvious.

Our simulation involved generating R = 4 choice responses to each of I = 5 items for N = 1000 persons. The item parameters were

$$(56)\qquad \begin{array}{ll} c(1) = 3.50 & c(4) = .50 \\ c(2) = 2.00 & c(5) = .286 \\ c(3) = 1.00 & \end{array}$$

The response parameters selected were

$$(57)\qquad \begin{array}{ll} \theta(1) = 1.000 & \theta(3) = -.3162 \\ \theta(2) = .3162 & \theta(4) = .0000 \end{array}$$

The alpha parameters were obtained by adding $|z|$ to a fixed constant for different persons. z is a random selection from a normal distribution N(0,1) and the constant was regularly modified so as to simulate persons with different levels of ability.

Thus the simulation program could compute R=4 odds according to (3) and hence R = 4 probabilities according to (7). These probabilities could then be accumulated and a uniform random number was generated and compared with the accumulated response probabilities. From this information a response was designated. This work was repeated I = 5 times for the same person and N = 1000 times for all persons in the sample. The response data and other information were punched on cards.

5.1.2 Data Tabulation

Assuming that the responses have been recorded on optical mark sense forms or have been placed on Hollerith cards we generate the frequencies in tabular form, $T(\substack{rs\\ij})$, shown in Figure 2. Notice the array of frequencies given in Figure 2 is doubly symmetrical. The array consists of blocks of responses, B(i,j) to pairs of items i and j. Within each block are the frequencies of response type to each item pair. For example, 63 persons responded in the second category to the first item and in the third category to the third item, etc. The total popularity of a response category is indicated in blocks B(i,i). For example, the third category of the first item is relatively unpopular.

The matrix $T(\substack{rs\\ij})$ is not required to be printed but its construction is required for subsequent phases of our analysis. The construction of this table concludes the Data Tabulation Phase.

5.1.3 c(i) Estimation

Corresponding to (22) for s = 1 and r = 4 we construct (54), $F(\substack{14\\ij})$, from $T(\substack{rs\\ij})$. This is easily done and produces

(58) $F(\substack{14\\ij}) =$

i \ j	1	2	3	4	5
1	-	83	126	163	179
2	49	-	96	135	117
3	41	44	-	94	84
4	18	27	43	-	55
5	14	22	26	35	-

		1				2				3				4				5			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	1	574	0	0	0	257	141	83	83	194	145	109	126	113	122	176	163	58	121	216	179
	2	0	206	0	0	87	51	35	33	[illegible]	49	63	44	30	59	55	62	22	46	90	48
	3	0	0	93	0	41	20	12	20	32	29	12	20	18	28	25	22	9	18	42	24
	4	0	0	0	127	49	29	20	29	41	30	27	29	18	30	40	39	14	30	49	34
2	1	267	87	41	49	444	0	0	0	161	101	86	96	78	113	118	135	47	103	177	117
	2	141	51	20	29	0	241	0	0	71	58	53	49	48	62	73	58	24	49	87	81
	3	83	35	12	20	0	0	150	0	41	40	34	35	26	25	55	44	10	26	69	45
	4	83	33	20	29	0	0	0	165	44	44	38	39	27	39	50	49	22	37	64	42
3	1	194	50	32	41	161	71	41	44	317	0	0	0	66	80	77	94	40	63	130	84
	2	145	49	29	30	101	68	40	44	0	253	0	0	44	71	74	64	20	59	98	76
	3	109	63	12	27	86	53	34	38	0	0	211	0	26	44	77	64	17	46	85	63
	4	126	44	20	29	96	49	35	39	0	0	0	219	43	44	68	64	26	47	84	62
4	1	113	30	18	18	78	48	26	27	66	44	26	43	179	0	0	0	16	41	67	55
	2	122	59	28	30	113	62	25	39	80	71	44	44	0	239	0	0	27	49	96	67
	3	176	55	25	40	118	73	55	50	77	74	77	68	0	0	296	0	25	63	124	84
	4	163	62	22	39	135	58	44	49	94	64	64	64	0	0	0	286	35	62	110	79
5	1	58	22	9	14	47	24	10	22	40	20	17	26	16	27	25	35	103	0	0	0
	2	121	46	18	30	103	49	26	37	63	59	46	47	41	49	63	62	0	215	0	0
	3	216	90	42	49	177	87	69	64	130	98	85	84	67	96	124	110	0	0	397	0
	4	179	48	24	34	117	81	45	42	84	76	63	62	55	67	84	79	0	0	0	285

Figure 2. Frequency table $T(^{rs}_{ij})$ for $F(i,j,k,L)$

According to (26) we next find the frequency ratios in R(i,j)

(59) R(i,j) =

i \ j	1	2	3	4	5
1		1.694	3.073	9.056	12.786
2	.590	-	2.182	5.000	5.318
3	.325	.458	-	2.186	3.231
4	.110	.200	.457	-	1.571
5	.028	.188	.310	.636	-

The skew-symmetric matrix resulting from taking logarithms of R(i,j) by (32) and (33) is

(60) L(i,j) =

i \ j	1	2	3	4	5
1	-	.527	1.124	2.203	2.548
2	-.527	-	.780	1.609	1.671
3	1.124	-.780	-	.782	1.173
4	-2.203	-1.609	-.782	-	.492
5	-2.548	-1.671	-1.173	-.492	-

Then according to (34), (37) and (39)

(61)

-T(i):	6.402	3.533	.051	-4.142	-5.844
t(i):	1.280	.707	.010	-.828	-1.169
$\hat{\epsilon}$(i):	3.598	2.027	.990	.437	.311
Item Parameters:	3.500	2.000	1.000	.500	.286

Notice that the item parameters and their estimates are in good agreement. This concludes the second phase of ε(i) estimation.

5.1.4 <u>θ(r) Estimation</u>

Our purpose now is to estimate the θ(r) from (40). To do this we use the procedures of section 5.1.3 to estimate the t(k) k = 1,4. This work has been completed and is summarized below:

Summary of the $t(\binom{k}{i})$ calculations for k = s = 2,3 and r = 4

(62)

i	k=2	k=3	a	b
1	.342	-.330	.2672	-.2578
2	.232	-.113	.3281	-.1598
3	.073	.014	7.3000	1.4000
4	-.284	.103	.3430	-.1244
5	-.362	.324	.3097	-.2772

According to (40) we estimate $\theta(2)$ and $\theta(3)$ by dividing column 2 and 3 entries above by the t(i) values obtained at the end of section 5.1.3. This division gives columns a and b above. According to our simulation columns a and b above show, embarrassingly, considerable variance from these values. Column averages, omitting i = 3 estimates give $\tilde{\theta}(2) = .3120$ and $\tilde{\theta}(3) = -0.2048$. Using these values the estimates of the response parameters are:

(63)
$$\theta(1) = 1.000 \qquad \tilde{\theta}(3) = -.205$$
$$\tilde{\theta}(2) = .312 \qquad \theta(4) = .000$$

Where agreement with (57), the parameters, is nothing to write home about even considering their questionable estimation procedure. This concludes the $\theta(r)$ estimation phase.

5.1.5 Goodness of Fit Test

Our purpose is to use (19) and our estimates of $\varepsilon(i)$ and $\theta(r)$ to compute the expected frequencies corresponding to (51). First, however, we obtain $G(\binom{14}{ij})$ as

(64) $(G(\binom{14}{ij}) =$

i \ j	1	2	3	4	5
1	-	132	167	181	193
2	132	-	140	162	139
3	167	140	-	137	110
4	181	162	137	-	90
5	193	139	110	90	-

For i>j use (20) for the lower segment of (65) and for j>i use (19) for the upper segment of (65).

(65) $P(^{14}_{ij}) =$	i \ j	1	2	3	4	5
	1	-	.6396	.7842	.8917	.9204
	2	.3604	-	.6719	.8226	.8670
	3	.2158	.3281	-	.6938	.7610
	4	.1083	.1774	.3062	-	.5842
	5	.0796	.1330	.1390	.4158	-

Multiplying (65) element-wise by (64) we get a matrix of expected frequencies $E(^{14}_{ij})$ as

(66) $E(^{14}_{ij}) =$	i \ j	1	2	3	4	5
	1	-	84.43	130.96	161.40	177.64
	2	47.57	-	94.07	133.26	120.51
	3	36.04	45.93	-	95.05	83.71
	4	19.60	28.74	41.95	-	52.58
	5	15.36	18.49	26.29	37.42	-

We calculate $\chi^2(i,j)$ using (58) and (66) with (67)

(67) $$\chi^2(i,j) = [F(i,j) - E(i,j)]^2 / E(i,j).$$

This gives

(68)	i \ j	1	2	3	4	5
	1	-	.0242	.1879	.0157	.0104
	2	.0430	-	.0396	.0227	.1022
	3	.6826	.0811	-	.0116	.0010
	4	.1306	.1053	.0263	-	.1114
	5	.0255	.6663	.0032	.1565	-

We next add the previous table symmetrically to give (69) and then add across the cols of (69) to give $\chi^2(i)$.

(69)	i \ j	1	2	3	4	5	$\chi^2(i)$
	1	-	.0672	.8705	.1463	.0359	1.1199
	2	.0672	-	.1207	.1280	.7685	1.0844
	3	.8705	.1207	-	.0379	.0042	1.0333
	4	.1463	.1280	.0379	-	.2679	.5801
	5	.0359	.7685	.0042	.2679	-	1.0765

The row totals are $\chi^2(i)$ with I-R degrees of freedom and measure the fit of each item to the model (7). The critical point for d.f. = 5-4 = 1 and α = .05 is 3.841. None of the $\chi^2(i)$ even approach this value, as expected, since the data were generated from the model's equation, (7). This concludes the goodness of fit phase.

5.1.6 $\omega(n)$ Estimation

See Section 3.5. This concludes the analysis for this example.

5.2 A Questionnaire Example

Following we provide a description of an interesting and useful situation to which the PMM and its analysis could be applied, but cannot be because the response data are hopelessly unobtainable from the NPRDC Magnetic Tape Library.

H. McDowell (1972) developed a series of questionnaires for each of several Naval ratings. The items of these forms were statements of tasks which must be performed by Naval Personnel on active duty who were employed in the rating. There were four responses for each task which indicated how independently, if at all, each task was performed by the Naval person. The responses to each task-item were:

a. I supervise the performance of this task.

b. I perform this task.

c. I assist in the performance of this task.

d. This task is not performed in my duties.

It seems to this author that the variable which these questionnaires could be thought to measure is "task-performance competency." For data which fit the model a measure of this variable could be constructed and applied to Naval personnel who are actively, or not, engaged in the measured rating. These measures would obviously be useful for advancement, assignment and re-assignment decisions.

Although the survey data were useful and informative for other purposes, they could have been of even greater benefit to the Naval Personnel System.

VI. SUMMARY

We have accomplished what was intended as stated in the introduction. Specifically, we defined the responses, relevant parameters and measurement situation related to a model called the Polychotomous Measurement Model. Both the theoretical and numerical analyses of the model were carried out in detail. Background material to, and applications of the model were provided.

VII. REFERENCES

Anderson, E. B. Conditional inference and models for measuring. Copenhagen: Mentalhygiejnisk Forlag, 1973.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29.

Fisher, R. A. Statistical methods for research workers (10th ed.). London: Oliver and Boyd, 1948.

Lord, F. M., & Novick, M. Statistical theories of mental test scores. New York: Addison Wesley, 1968)

McDowell, H. New concepts to improve training and shipboard manpower utilization in the engineering and hull ratings (NPRDC Report SRR 72-20). San Diego: Naval Personnel Research and Development Laboratory, 1972.

Moonan, W. J. Contributions to objective measurement and evaluation of trainee competency. Proceedings of the 1972 Military Testing Association Conference, Oklahoma City, 1972.

Rasch, G. Probabilistic models for some intelligence and attainment test. Copenhagen: Danmarks Paedogogiske Institut, 1960.

Rasch, G. On general laws and the meaning of measurement in psychology. Proceedings of the 4th Berkeley Symposium on Mathematical Statistics, IV, Berkeley, CA., 1961.

Vogt, D. An extension of the Rasch Model to the case of polychotomously scored items. Ph.D Dissertation, University of Maryland, College Park, MD., 1971.

Wright. B. D. Sample free test calibration and person measurement. Proceedings of 1967 Invitational Conference on Testing Problems, Educational Testing Service, Princeton, N. J., 1968.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23.

ENLISTED ARMY JOB EVALUATION METHODOLOGY AND RELATED POLICY ALTERNATIVES

Raymond O. Waldkoetter, Ed.D.

U.S. Army Research Institute for the Behavioral and Social Sciences[1]
Fort Sill Field Unit
Fort Sill, Oklahoma 73503

INTRODUCTION

To develop a force structure or even a small organizational structure there is ordinarily a need to align duties by grade. This is a simple matter of distinguishing responsibility and providing the best gradation of skill and expected performance. As the Department of Army faces changing personnel requirements in organizational design for particular missions and units and for comprehensive restructuring such as the Enlisted Personnel Management System (EPMS) changes, there must be available a reliable and valid methodology to determine objective standards of grade authorization.

Since the pay grade allocation for enlisted jobs (E-1 through E-9) generally carries increasing levels of responsibility and opportunity, the job requirement factors require accurate estimation and scaling to develop pertinent grade standards. Job factors such as Knowledge and Combat Exposure are rated to identify the level of the factor needed to expect a job will have sufficient skill and authority invested to get the desired productivity.

The DA personnel analysts and various MOS job proponents and major commanders have to have a methodology which can dependably identify grade standards and which will relate to the total enlisted job structure. This methodology has to compute grade optimally, but occasionally even must work when grade is constrained to accommodate curtailed grade allocation to satisfy grade ceilings.

---

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institute, the Military Personnel Center, or the Department of the Army.

Because grade determination follows a given administrative and psychometric procedure, the methodology becomes a combination of judgmental and rating behaviors. With the grade assignment process managed at the DA level and recommendations for new positions or grade changes coming into the designated action personnel, there is quite often a tendency for the requesting grade adjustment to appear out-of-balance on at least one job factor or grade standard. The grade estimate then must translate what appears at times to be rather subjective judgment into objective and predictable grade.

For many years the grading of jobs or positions could rely on traditional military structure and in most cases this reference to experience and known mission requirements proved sufficiently adequate. From World War II the Army began carefully documenting its techniques, policies, and guidelines for assigning grade to enlisted jobs (Hadley, 1961). It became more and more apparent that an approach that was quantifiable and statistically defensible would have to form the foundation of the Army job evaluation methodology. Such methodology would help in distinguishing the job requirements most equitably while supporting the best allocation of grade for job types and functions to permit better organizational and grade balance projections for force and budgetary planning. At this point past approaches were reviewed, and a concerted effort was directed to explore the predictive value of possible enlisted multiple regression equations along lines already pursued experimentally for Air Force officer positions (Christal, 1965).

In 1966-67 the traditional and accepted job factors were used to develop several mutiple regression equations to explain how enlisted grade could be derived systematically and to explore the policy agreement of selected raters evaluating job factors and the associated grade (Anderson, Corts, and Waldkoetter, 1967). This effort was modestly successful and provided a pragmatic beginning for enlisted job evaluation in that the sample of jobs rated and the policy-board grades assigned only responded to a request for an equation model. Treating any implications for those situations arising with application of an equation, which is then tested against other policy constraints, was simply avoided.

After nearly eight years and numerous personnel policy adjustments leading to doctrine, training and force structure changes, the question about equation policy implications was raised in the form of a research need in accordance with the Army Regulation 70-8 (1975). In 1975 a few key policy changes regarding the EPMS, the objective force grade constraints, and the career force development led to the situation of asking whether the job factors, the related formula weights, and implications of any equation usage were really functioning in the same original context of 1967. Knowing that grade is compensation for money and work identity and that economic indexes had shifted significantly in the post-Vietnam era, it was believed important to reassess the basis for the 1967 equation and account for any noticeable reactions to changing standards for grade determination.

METHOD

A sample of 200 enlisted benchmark jobs or duty positions was selected. A group of military personnel analysts slightly modified the sample content toward more combat arms positions to retain a concept that any job grading procedure should have an orientation lending more influence to the basic mission of the force. These analysts were aware that the initial sample was constructed by a stratified random plan of selection. Once the sample was agreed on by the analysts from the U.S. Army Military Personnel Center preparations were made by the same analysts to write the 200 duty position job descriptions, which were supported using a bank of CODAP duty position descriptions and others maintained in their usual processing of job actions.

Concurrently, the 10 job factors used for the 1967 equation were reviewed to see that the definitions and scale values from one through six were still acceptable for rating purposes, with two experimental job evaluation or requirement scales being newly defined. The scales of Job Satisfaction and Organizational Setting were designed to assist in getting at evaluation actions where a job's worth is difficult to estimate: first, because job satisfaction within a job may exist as an inherent grade bias toward higher grade or satisfaction decreases with grade when grade should offer a career attraction; and secondly, because organizational setting is positively weighted regardless of the type of grade or job evaluation action, may contribute too much toward some grade allocation, and had not been quantitatively analyzed. The research design then called for using 12 job factor scales as predictors of grades E-3 through E-9, the scales are: (1) Knowledge; (2) Supervision; (3) Concentration and Attention; (4) Freedom of Action; (5) Physical Effort; (6) Combat Exposure; (7) Adaptability and Resourcefulness; (8) Responsibility for Material Resources; (9) Physical Skills; (10) Job Conditions; (11) Job Satisfaction; (12) Organizational Setting.

Three groups of policy boards were identified to estimate grade for the duty positions and were comprised of 50 MILPERCEN officers, 75 Fort Harrison officers, and 75 Fort Bliss NCO's. The three boards also independently rated 12 job factors. Policy board officers and NCO's each rated 20 duty position descriptions with all 200 benchmark job descriptions rated in each group by means of 10 stratified sets of 20 randomized job descriptions. Thus several independent situations were arranged with equivalent rating scenarios to determine the degree job factors were required to perform the duties for each job and the most equitable grade. The choice of three groups provided for several comparative criterion policy building steps. Also the range of rating behavior would likely vary sufficiently to indicate to some extent if the prediction of grade would stabilize enough to warrant use of any given equation within an operational set of limitations related to the career structure, work design, and duty position identity. The job sample size of 200 benchmark positions and selection of policy groups were believed adequate to obtain sufficient prediction of grade that reasonable job factor reliability

would result and a grading equation would function quite credibly through the different levels of personnel review and accounting.

Several policy choices were developed exploring what would happen when the 200 officers and NCO's and combinations thereof furnished job factor ratings and grade judgments. The MILPERCEN and Fort Harrison officer factor ratings were used to predict criterion grade ratings of the Fort Bliss NCO's. Then the NCO factor ratings were combined separately with each officer group to predict each set of officer criterion grade ratings. This would allow for a range of possible factor ratings to observe the effects on grade prediction while identifying the combination yielding the highest multiple correlation coefficient. Such an approach also gave a notion of equation differences which might arise as factor weights are varied and criterion grade ratings are alternately substituted. Although the validation of the best equation rests on comparison of the three equations, a certain fundamental stability of the highest multiple correlation coefficient was being confirmed by obtaining correlations which would show they were relatively close together. In this instance it is surely a stringent alternative procedure for cross-validation because there are deliberate variations introduced rather than randomly dividing the groups into two subsamples and having the same criterion grade judgments used to compare equations.

Policy choices further considered that equations represented different perceptions of what might serve to determine grade. Should a mix of officer and enlisted job factor ratings offer a more desirable grade prediction? Or would the use of officer factor ratings best predict enlisted criterion grade ratings? Then, too, could fewer factors predict grade almost as accurately as all factors? These questions occur since they can influence how job grading is done and whether particular values are consistent in the grading process.

Besides trying to derive the optimal grading equation, there was a persistent question about the vulnerability of a job to a downgrade action if a requirement is directed to lower grade or hold to given grade constraints to keep the force within specified boundaries. The process of grading jobs poses a series of conflicting requirements at times to provide the most equitable grade for the job and to hold the job structure to authorized force levels or budgetary costs. Any policy should develop according to specified steps. More often than not the best grade prediction should act as a solid reference point and come from an independent rating process, then respond to the limits imposed by personnel structure and grade quotas. A questionnaire was administered to check on the rating effects of rater opinions and what might help in the hopefully objective process of job grading and pay allocation systems.

RESULTS

After the 200 benchmark job descriptions were rated on the 12 job requirement factors with proposed grade, the data were reduced to have estimates of

factor reliability and stepwise multiple correlations with factor validity coefficients computed. The reliability estimates using analysis for variance (Winer, 1962) were computed for each job factor to assure that an adequate level of reliability was secured in the collected ratings. Reliability coefficients ranged from .87 to .98 with most falling in the middle ninety range. The reliability coefficient for the mean grade ratings across sampled jobs was .98. For the experimental scales, Job Satisfaction and Organizational Setting, the reliabilities were .93 and .96, respectively. Validity coefficients for job factor ratings correlated with the criterion of proposed grade for the combined groups were generally high to very high ranging from .56 to .90, but across the three combined groups of policy board raters four factors had negative or extremely low validity. These were Physical Effort, Combat Exposure, Physical Skills, and Job Conditions. Both Job Satisfaction and Organizational Setting indicated high validity (.82 and .88).

Multiple regression (R) equations were derived for three combined groupings: MILPERCEN and Harrison with Bliss grade criterion; MILPERCEN and Bliss with Harrison criterion; and Harrison and Bliss with MILPERCEN criterion. The multiple R's were derived producing .94 using the Bliss criterion grades, .93 using the Harrison criterion, and .92 using the MILPERCEN criterion. Although the regression coefficients used to multiply each factor weight varied to a greater or lesser extent in arriving at the multiple R for each combined group, the end result was a highly predictive equation. The equations almost uniformly predicted current or actual grade given for the 200 job descriptions at .83. Perhaps rather oddly the 1967 equation predicted actual grade then with a multiple R of .83. Superficially, none of the three equations predicting optimal grade seems to give much advantage in predicting current grade, yet an observation is suggested that the intervening 10 years and use of the 1967 grade equation under personnel management constraints did not materially distort the grade allocation order.

Since the multiple R of .94 was derived from the MILPERCEN and Fort Harrison officers factors ratings correlated with the Fort Bliss NCO's criterion grade ratings, it was tentatively agreed that this equation would provide more probable credibility. The officer factor ratings suggest that officers must capably direct job requirement standards, while the NCO's should know most competently what an enlisted job is comparatively worth and must obtain corresponding productivity. More analyses are being performed to make sure this equation has all of the statistical characteristics to recommend a prototype application for a series of job grading requests. There have already been some trial data collected which give a trend of lower grade estimates than obtained with the 1967 equation, but the sample of jobs is not representative and the personnel analysts were not too experienced using the new equation. As it stands the proposed constant value and weights for mean factor ratings in the new equation are:

$$1.343 + 0.541\overline{X}_{I} + 0.233\overline{X}_{II} + (-0.353)\overline{X}_{III} + 0.258\overline{X}_{IV} + (-0.087)\overline{X}_{V} + 0.039\overline{X}_{VI}$$
$$+0.398\overline{X}_{VII} + 0.093\overline{X}_{VIII} + (-0.016)\overline{X}_{IX} + 0.046\overline{X}_{X} + 0.013\overline{X}_{XI} + 0.132\overline{X}_{XII}$$
= computed grade.

When the first (lowest) level for all 12 job factors (i.e., $\overline{X}_{I}, \overline{X}_{II} \ldots \overline{X}_{XII}$) is entered into this equation, the result is 2.63. This value would probably round to E-3. A shortened version of the factor equation in the stepwise regression gave the constant value and positive weights of:

$$1.298 + 0.408\overline{X}_{I} + 0.252\overline{X}_{II} + 0.242\overline{X}_{IV} + 0.254\overline{X}_{VII} + 0.187\overline{X}_{XII} = \text{computed grade.}$$

When the first (lowest) level for all five factors is entered into this equation the result is 2.64. This equation is a practical substitute when a quick estimate is asked for and the 12 factor equation is applied later for review purposes.

The downgrading equation, or method by which downgrading could occur, examined the policy of raters who would perform downgrading of an available sample of enlisted positions. The rating groups were directed to reduce 10 of 20 job descriptions by one grade. The equation, then, predicted if a job would have the factors for downgrading. Significant agreement seemed related to three job factors with a negative value for Organizational Setting (XII), and positive, counter-balancing values for Combat Exposure (VI) and Responsibility for Material Resources (VIII). The equation having a large constant of 9.38, then had the following weights $-1.30\overline{X}_{XII} + .70\overline{X}_{VI} + .80\overline{X}_{VIII}$ derived which are used basically to give an estimate for probability of a "non-downgrading" action. The objective of predicting 10 downgrading decisions by 20 raters, led to the computation of a low-value index by dividing the above equation solution by 20 to determine the non-downgrading probability, with the probability of downgrading becoming greater as the numerator becomes smaller. This approach only illustrates how downgrading may have systematic formulation to guide such policy usage and control. Pragmatically, it was found the factor weights could be substituted in the 12 factor equation and the result was a realistic operational estimate of whether the job might appear vulnerable to downgrading. An example of the projected use of the downgrading equation would result in the probability of downgrading a position like Reenlistment NCO but the job of Rifleman would retain the assigned grade.

A job rating questionnaire (JRQ) was given to all members of the three policy rating boards. This questionnaire was used to obtain possibly relevant information connected with downgrading and rating behavior, interpretation of grade prediction equations and certain biodata. The interpretive sections of the questionnaire explored three areas which could systematically influence

policy outcomes: job factor methodology and procedures; career management considerations; and economic attitudes. The results suggest that the job factors were adequately defined for the rating task and that the raters brought their own judgment to bear with systematic grading standards. Responses indicated the majority of judges found the technical areas and special skill positions most difficult to evaluate. This finding may suggest that a vertical grade structure is not the best way to reward technical-skill performance, and that level of authority is hard to balance with amount of specialization. An implication may encourage further research toward the development of a skill progression and reward system separate from grade progression. In terms of career management a variety of considerations were dealt with relevant for promotion and regrading actions. It was observed that nearly 80 percent of rater-judges believed that current policies have resulted in the grade structure and promotions being more competitive than five years ago. There was fundamental consensus among the judges concerning retirement promotion with 66 percent indicating promotion should be contingent upon evidence of at least one or more years of active duty remaining before retirement. Given a choice between down-grading some jobs or holding grade for all jobs in order to meet budgetary constraints, 26 percent of the judges chose the former and 36 percent chose the latter with the others voting to analyze the problem from a different approach. The results may tend to suggest that interfering with career progression is not viewed positively to satisfy budgetary constraints. In exploring the economic attitudes of the judges, 66 percent indicated they believed their grade estimates provide the income and recognition required to assure adequate job performance.

Further, it was noted the promotion concept is not easily compensated for and any approach to offer other types of rewards and recognition will somehow have to be related to some concept of continued advancement or career progression. Another observation occurred to imply that any personnel-training actions tied to promotion delay or substitution will not be readily accepted unless promotion is treated as a separate transaction and any cost problem is adjusted to show it is not managed to affect promotion. The background information (biodata) qualified the policy raters as having 75 percent with 10 or more years of service, working in a wide cross-section of units, the majority having three years or less time-in-grade, 87 percent educated beyond the high school level, and 94 percent having training for their military duties through resident courses and OJT.

DISCUSSION

The Army Research Institute in conjunction with the Military Occupational Development Division, MILPERCEN has explored a new validation and adaptation of job evaluation methodology in developing tentatively proposed enlisted grade equations. The job requirement factor ratings given by 125 officers (Captain through Colonel) accurately predict the grade assigned by a policy board of 75 non-commissioned officers (E8/9) selected for

experience and their extensive knowledge of enlisted duty positions. The principal equation developed predicts enlisted grade by making use of the strong relationship between quantitative estimates of 12 highly reliable and generally valid job factors and the proposed grade for 200 well selected "benchmark" duty positions.

The 12 factor and 5 factor equations provide for contemporary improvement beyond the 1967 equation and related job evaluation materials. The job evaluation factors were the same for 1977 with the exception that two factors, Job Satisfaction and Organizational Setting were added, received positive weights, and Organizational Setting proved a very strong predictor. From 1967 to 1975 before the recently completed research, minor editing and improvements were made on the original 10 job factors which were defined along traditional lines to be compatible with other DOD approaches and Civil Service to express job duties in terms of similar, logically derived scales of job requirements.

Job factor methodology as presented by the 1977 equations provides a reasonably valid grading and decision-making process. Similar factor systems have been developed by and updated by the Civil Service Commission (Anderson and Corts, 1973) and Air Force (Christal, 1975). The quadrennial review of military compensation (Pappas, Fisher, and Martin, 1976) sought as a study objective to link certain military jobs with their civilian equivalent and did not appear to find any other system more valid than the job factor equation. Some direct use of task inventories may offer supplemental data for assigning grade with a basis for determining the grades of personnel performing similar sets of tasks. The consensus of expert judgment which gives validity to the job factor methodology is based essentially on a "voting process," so that any later adjustment to a job grade results likely from variables such as MOS grade structure or some other grading guidelines. Comparative analyses of the Armed Services and wage scale linkage in the civilian economy, or even grading policies of major corporations and labor unions, can provide more insight regarding the relationships of grade, wage scale, and overall compensation. No matter which line of grade estimating you follow, the job factor approach will provide an accurate base estimate from which grade determination can proceed so that analysts are better oriented to apply required policy guidance.

The downgrading equation was not highly significant yet points the way toward designing a unique policy strategy for reacting to grade constraints or limitations. This procedure along with a firm "rounding" guideline when averaging position grade estimates gives a more conservative and slightly lower grade projection that is arbitrated by expert analyst judgment using at least five independent job ratings and Army policy guidance.

The Job Rating Questionnaire is meant to offer certain supplemental information related to grading and MOS structure design. This information can help in modifying selected personnel actions so that the career patterns and MOS management are more related to the policy understanding of officer

and enlisted personnel who must perform the operational functions of personnel utilization. The strength and relatively obvious interpretation of equation and questionnaire results did not recommend further research to differentiate possible effects of rater characteristics. However, after continuing purification of the data base, a multiple R of over .97 has resulted and some explanation of rater effects will follow in a later report.

In summary, the job evaluation methodology was refined and is being revised by extended Army Research Institute effort to be flexible in relation to externally imposed grade constraints. Policy forecasting questions were explored for further development so that grade requirements can adjust to changing policy for time-in-grade, MOS strength, and grade-pay phasing. Career progression design can have a more equitable basis and future Army enlisted grading can apply highly reliable standards when deciding job value.

REFERENCES

Anderson, C.H. and Corts, D.B. Development of a framework for a factor-ranking benchmark system of job evaluation. TS-73-3. Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, December 1973.

Anderson, D.H., Corts, D.B., and Weldkoetter, R.O. An operational model for the development of an optimal US Army enlisted grade structure through job evaluation. EEC-RS-100, AD-660327. Indianapolis, Indiana: US Army Enlisted Evaluation Center, October 1967.

Army Regulation 70-8, Human Resources Research Program. Washington, D.C.; Headquarters, Department of the Army, 13 March 1975.

Christal, R.E. Officer grade requirements project: I. Overview. PRL-TR-65-15, AD-622806. Lackland AFB, Tex: Personnel Research Laboratory, Aerospace Medical Division, September 1965.

Christal, R.E. Systematic method for establishing officer grade requirements based upon job demands. AFHRL-TR-75-36, AD-A015 756. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, July 1975.

Hadley, H.I. The current system of Army job evaluation. Paper read at Tri-Service Conference on the role of job evaluation techniques in the structuring of military occupations. ONR Symposium Report ACR-71. Washington, D.C.: Department of the Navy, April 1961.

Pappas, L.D., Fisher, Jr., A.H., and Martin, Jr., F.B. An analysis of selected linkages between military and civil service occupations. TR-76-2. Washington, D.C.: Hay Associates, April 1976.

Winer, B.J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

DETERMINATION OF AIR FORCE NON-AIRCREW OFFICER GRADE REQUIREMENTS

By

Kenn Finstuen
and
Lt Col William H. Pope, USAF

Occupation and Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Brooks AFB, Tx 78235

Prepared for symposium:
Job Evaluation in the Military, Government, and Industry

19th Annual Conference of the Military Testing Association
17 - 19 October 1977
El Tropicano Hotel
San Antonio, Texas

I. INTRODUCTION

The Officer Grade Requirements (OGR) research project was designed to assess the appropriate grade (lieutenant through colonel) for each of the 62,000 jobs in the total non-aircrew officer force. The job evaluation technology which was developed in the course of the OGR project is a systematic and reliable method which can be used to determine grades for non-aircrew jobs based on job content and responsibility. The present updated version of the technology was developed for use by the Air Force Management Engineering Teams (METs) to be employed as a manpower tool in assigning grades to jobs.

For those not familiar with the Air Force occupational structure, there are 54 career fields or job groupings for officers (excluding pilots and physicians) which are made up from over 180 Air Force Specialties. As an example, Table 1 reflects the scientific career field made up of seven specialties.

Table 1. Example of Air Force Job Structure

One Career Field	Comprised of:	Seven Air Force Specialties	
Scientific		2616	Scientific Manager
		2635	Physicist
26XX		2645	Chemist/Biologist
		2655	Metallurgist
		2665	Nuclear Research Officer
		2676	Behavioral Scientist
		2685	Scientific Analyst

II. REVIEW OF OFFICER GRADE REQUIREMENTS RESEARCH

The basic technology for the OGR job evaluation method has been under development and refinement for over 15 years. To understand the methods and procedures used in the present study a brief explanation of the initial phase of the OGR (1963-1966) is presented below, prior to a discussion of the latest job evaluation exercise employed by METs.

Six steps were involved in the first phase of the OGR project. Figure 1 depicts actions, dates and job sample sizes for these six steps.

Step 1. Collection of Job Descriptions (1963)

The first step of the project involved collection of accurate and detailed job statements describing the work performed by Air Force officers in all Air Force specialties. Figure 1 shows the initial 1963 job collection of 79,750 officer jobs. Each job incumbent provided a job title, a description of his job in the Air Force organizational structure, and a detailed description of duties and tasks performed.

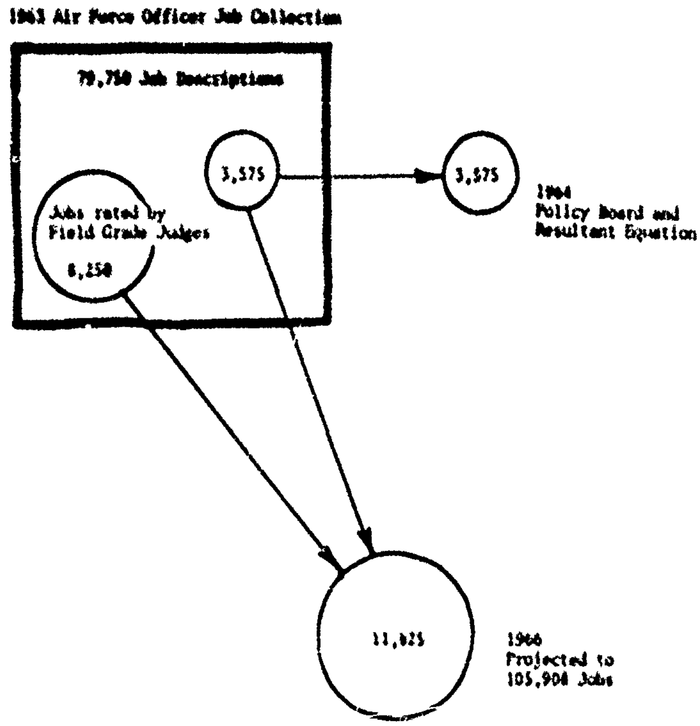


Figure 1. 1963 - 1966 Officer Grade Requirements

Job Evaluation Development

The incumbent's supervisor was asked to review the job description for accuracy and provide a judgment concerning the appropriate grade for that job.

Step 2. Selection of Criterion Sample (1963-1964)

The second step consisted of selection of 3,575 representative jobs in all grades across all major air commands, both overseas and in the continental United States. Jobs were selected from all Air Force Specialty Codes (AFSCs).

Step 3. USAF Policy Board Ratings (1964)

In the third step, a policy board composed of 22 colonels was convened by HQ USAF. Policy board members were selected on the basis of their experience in various career areas so that for any of the 3,575 criterion sample jobs, there was at least one member who could serve as an expert consultant to the rest of the board in making judgments as to the appropriate grade (lieutenant through colonel) for the jobs. Board members were instructed to judge each job on its merits.

In quantifying and recording the board's judgments the following measures were taken:

a) Board members rated the appropriate grade level for a job and then indicated on a 3-point scale their level of confidence in such ratings. They were given access to any information needed to make accurate judgments. This included consultation with other members, obtaining organizational, command, or installation information about a job, and telephoning special air staff consultants or the supervisor of the incumbent of the job being rated. However, members were advised that their ratings were to be independent and were to reflect the unbiased judgment of the rater alone. The board members did not have knowledge of the authorized (Unit Manning Document-UMD) grade for the job being rated nor of the grade stated by the incumbent's supervisor. They were not informed of the grade held by the incumbent in the job, nor were grade ratings assigned by other board members available to the rater.

b) Grade ratings for particular jobs were obtained independently from five separate board members since earlier Air Force job evaluation research indicated that an average of five independent ratings provided stable estimates for job evaluation purposes.

c) Each job was rated in a context of other jobs since earlier job evaluation research on context effects suggested that more accurate ratings of job level are obtained when a job is considered with other jobs of varying content and level.

d) Board members rated grade requirements using a 16-point rating scale consisting of three levels of experience within each grade from lieutenant through colonel, and one level for general. This scale was based on findings that ratings are more stable when judges made the finest discriminations of which they are capable, and the assumption that experienced officers can distinguish jobs requiring high, moderate, or low levels of experience.

Step 4. Analysis of Policy Board Actions (1964)

Analysis of the policy board rating data was a critical part of the project since these ratings formed the basis for establishing grade requirements. A series of analyses was accomplished to determine if the grade ratings were stable; if there was high agreement among board members concerning the appropriate grade requirements for particular jobs; if the raters had confidence in their ratings; and if the raters were biased for or against jobs in various AFSCs or commands. Main results from these analyses were as follows.

a) The reliability coefficient (.92) of the mean grade ratings as given by the policy board indicated there was high agreement among board members concerning grade requirements for jobs in the criterion sample.

b) Based on a 3-point scale of confidence, (1 = low to 3 = high) board members expressed a high level of confidence in ratings with at least 4 or 5 raters expressing the highest level of confidence in their ratings of 2,389 of the 3,575 jobs. In fact, only 59 of the jobs had a confidence rating of less than 2.

c) Analysis designed to identify raters showing a bias for or against jobs by command or occupational grouping revealed that individual board members did not exhibit a bias towards jobs in any particular command or AFSC. The highest reported disagreement among board members varied from the mean by only 1.7 points on a scale from 1 to 16, with most values being less than 1. The maximum level of disagreement among raters varied by less than 1/2 of a grade.

d) Additional analyses indicated that board members agreed that many jobs were inappropriately graded and that each job was considered on its own merits. Comparison of UMD versus board grade revealed no systematic tendency on the part of raters to confirm current UMD grade authorizations or to inflate their ratings of grade requirements. Many jobs were downgraded as much as one or two full grade levels, while others were upgraded. Reliability analyses indicated there was strong agreement among board members as to which particular jobs should be upgraded and downgraded.

Step 5. Development of Policy Equation (1964)

The fifth step of the initial OGR development entailed the development of a multiple linear regression equation which "captured" the policy-making grade decisions of the criterion board using several variables with judgments obtained from 1,246 field grade raters on the 3,575 job set. A high agreement among board members concerning the appropriate grade for jobs indicated they considered similar factors in making their decisions. In developing the equation, over 200 predictor variables and 350 regression problems were considered. As a result of these analyses, predictors were eventually narrowed down to 9 variables involving 1) special training and work experience, 2) communication skills, 3) judgment and decision making, 4) planning, 5) management, 6 & 7) two levels of organization information, 8) the supervisor's judgment of grade for the job (lieutenant through general), and 9) the field grade judge rating.

Step 6. Application of Policy Equation (1965)

The last or sixth step in this initial phase of job evaluation developed consisted of applying the grade policy equation to an additional 8,250 job descriptions rated by field grade raters (see Figure 1). Together with the 3,575 policy board ratings from jobs, appropriate grades for over 11,000 officer jobs had been determined. From this large base it was possible to project to the grade structure of the entire Air Force consisting of 105,908 jobs, and to compare projected grade requirement statements with the total Air Force authorizations (UMD).

Results indicated that changes were necessary in many career fields in order to align grades with jobs based on job content and responsibility levels as stated by OGR. For some career fields such as Research and Development, Developmental Engineering and Financial the OGR study indicated that too few colonel positions were available; while in other fields such as Education and Training, Information, and Legal, too many colonel positions were allocated. Overall, many jobs were found to be undergraded, especially at the rank of major where greater content and responsibility levels existed than what UMD authorizations indicated.

<u>Development of Bench Mark Rating Scales</u>

In addition to the above technology, in 1966 benchmark scales were developed for the OGR job evaluation factors of special training and work experience, communication skills, judgment and decision making, planning, and management. Scales range from 1 to 9 with three job titles chosen to represent each of the nine increasing skill levels. These scales allow more consistent and reliable judgments to be made by comparing isolated or single jobs with the benchmarked jobs of the scale. The scales were validated using 1,000 officer positions. The job grades obtained with benchmark scales compared with the Policy Board were in high agreement (r = .90). An example of a current benchmark scale for the management factor is presented in Table 2.

Table 2. An example of a benchmark scale.

MANAGEMENT: The level of executive, and managerial skills required in the job. Consider the complexity, variety and level of the activities which are directed, organized, coordinated, controlled, commanded, or evaluated

LEVEL 9

Director of Budget, Hq Major Air Command
Commander, Combat Support Gp (Overseas)
Wing Commander, Tactical Control Wg (Overseas)

•
•
•
•

LEVEL 7

Administrative Officer, Air Base Sq
Data Services Officer, Combat Support Gp
Tactical Fighter Pilot, Tactical Fighter Sq

LEVEL 1

Clinical Psychologist, USAF Hospital
Psychiatric Social Worker, USAF Hospital
Helicopter Pilot Single Rotor, Air Base Sq

III. DEVELOPMENT OF JOB EVALUATION SYSTEM FOR MET's

The first development of the technology using MET personnel began in 1974. Previously the technology had employed field grade officers as raters. The 1974 pilot test confirmed that METs could also apply the technology with accuracy and consistency.

The process used in the determination of grades is presented in Figure 2. The basic process involves seven actions from the time the job description is filled out until the appropriate grade for the job is assigned. After the supervisor makes a grade judgment, MET raters make judgments on five grade factors with the aid of benchmark scales which were discussed previously. When this information is combined with organization information, resulting scores can be converted to a grade which reflects the job content and responsibility level of the job.

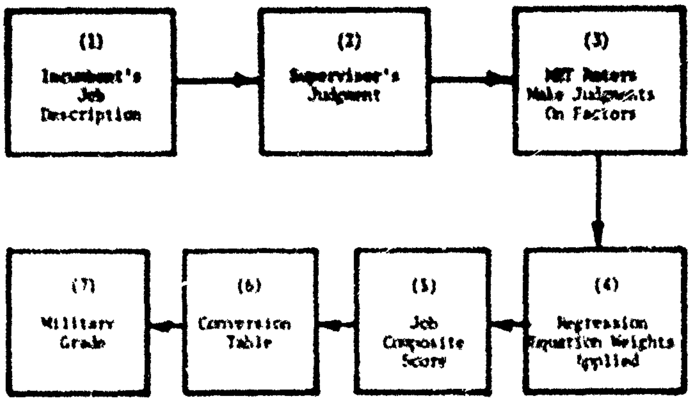


Figure 2. The grade determination process.

Job Samples

Job descriptions were collected by MET members and rated on job factors. In the 1974 pilot test, 1,687 job descriptions were collected to test the feasibility of MET application of the OGR system. In addition, 485 of the original 3,575 criterion board jobs were rated by MET members to assess the validity of the OGR technique. Reading the crossbreak horizontally in Table 3 shows 2,172 jobs rated in the 1974 pilot study, consisting of 485 policy board jobs and 1,687 current Air Force officer jobs which were collected by METs. Successful completion of the pilot test indicated that METs could systematically and reliably determine grade requirements for jobs, and a full field test was therefore undertaken to refine and expand this job evaluation technology. Samples from the 1974 pilot study were incorporated into the crossbreak design of the larger effort as shown in Table 3. In the 1976 study, as in the 1974 pilot study, two kinds of jobs were rated. One set consisted of jobs from the original 1964 policy board set and the other consisted of current Air Force job descriptions collected during the study.

Reading the crossbreak vertically, rather than horizontally, it may be seen that, of the policy board jobs, a total of 1,725 criterion jobs were rated by MET members. In addition, over 11,000 new jobs were collected and rated in this phase of the OGR. This 11,000 case sample was used as the base from which projections were made to the total non-aircrew force of over 62,000 Air Force officer positions. In all, METs evaluated over 13,000 jobs in this most recent undertaking of the OGR research project.

Table 3. Sample Crossbreak of Subsets and Total Number of Jobs Used In Job Evaluation

	Policy Board Jobs	Current A.F. Jobs	Totals
1974 Pilot Study	N = 485 (1)	N = 1,687 (4)	N = 2,172 (7)
1976 Main Study	N = 1,240 (2)	N = 9,634 (5)	N = 10,874 (8)
Total Sample	N = 1,725 (3)	N = 11,321 (6)	N = 13,046 (9)

Cells	Description
(1) to (3)	Policy Board jobs used as the criterion in computing validity and construction of the grade conversion table
(4) to (6)	Current jobs newly collected used as the base for making projections to the total non-aircrew force
(7) and (8)	Rating subset information used for reliability computations
(9)	Grand total number of all jobs used in the study

The 1,725 criterion sample set of jobs consisted of Air Force non-aircrew jobs which are still in the Air Force today. The sample was used to derive a new set of regression equation weights for the Policy equation and to construct a conversion table for converting composite scores to military grades.

The second, current job sample, was used to project grade requirements to the total non-aircrew force. Sampling specifications were established for the 122 METs, with jobs stratified across Air Force specialties and across grades (lieutenant through colonel) in order to assure that descriptions were representative of the total non-aircrew force. Sample specifications were based on a December Unit Detail Listing (UDL - present version of the UMD). Over 900 MET raters from all over the world participated in the study.

Analyses of 1,725 Criterion Jobs

Analyses resulted in an 8 variable multiple regression equation using the five factors, two level of organization variables and the supervisor's judgment. When compared to the grades assigned by the Policy Board to the same jobs the Policy Equation reflects a validity coefficient of .90. Individual validities for each of the variables are reported in Table 4. Based on these findings, the 1,725 job set was used to construct a stable conversion table to convert the composite scores obtained from the application of the regression equation to a military grade. This permits the evaluation of a job, using the grade equation to provide a score, and a conversion of the resulting score into the appropriate officer grade required for the job.

Table 4. Validity of Grade Predictors In The Grade Equation for the OGR

Variable	Validity
Special training and work experience	.65
Communication skills	.72
Judgment and decision making	.74
Planning	.78
Management	.79
Level of organization	.50
Level of job within organization	.47
Supervisor's judgment of appropriate grade	.78
Final grade composite score	.90

Based on an average of 13.61 ratings per job

By using the grade equation, this method will assign the same grade to each non-aircrew position as would have been assigned by the Policy Board members. Table 5 indicates the variables of the equation and their associated weights used in determining a composite score. The mean values of 5 job evaluation factors rated on benchmark scales are weighted and combined with the two organization variables and the supervisor's judgment.

Table 5. Application of the Policy Equation

GRADE DETERMINANTS	JOB VALUE	WEIGHT	WEIGHTED COMPONENT
Special training and work experience	$\bar{X}$	1	#
Communication	$\bar{X}$	1	#
Judgment and decision making	$\bar{X}$	2	#
Planning	$\bar{X}$	1	#
Management	$\bar{X}$	3	#
Level of organization	JD*	1	#
Level of job within organization	JD*	1	#
Supervisor's judgment of appropriate grade	JD*	**	#
			COMPOSITE SCORE

* from job description
** + or - depending on grade colonel through lieutenant

Analyses of 11,321 Current Jobs

To examine how well the MET raters agreed among themselves as to a jobs content and responsibility, inter-rater reliability coefficients were computed. Based on 6.94 ratings per job, the intra-class coefficient resulted in $r_{kk} = .97$. This coefficient indicated very high agreement among the raters and indicated that if the 11,000 jobs were evaluated by another group of raters, similar results would be obtained.

IV. PROJECTIONS TO THE TOTAL NON-AIRCREW OFFICER FORCE

Based on the preceding findings and indications, it was possible to project, with considerable accuracy, the results which would be obtained if the METs were to apply OGR technology to the entire non-aircrew force. The 11,321 sample was representative and large enough (18 percent of the force) to provide stable projections of the various career utilization fields for the 62,602 positions listed on the December 1975 Manpower and Organization UDL (M&O UDL).

Table 6 presents the overall results from this projection exercise, comparing the authorized manpower statements of the UDL, OGR statements of grade, and the December 1975 actual on-board manning strength as reflected in the Uniform Officer Records (UOR). As may be seen, OGR calls for a decrease in the number of currently authorized colonels, an increase in lieutenant colonels, a very large increase in majors, and a significant decrease in the authorizations for captains and lieutenants. In summary, an overall increase in the number of field grade officers would be needed to match the grades of the incumbents with the content and responsibility levels of Air Force non-aircrew officer jobs.

Table 6. Projections to the Total Non-Aircrew Force

Grade	M&O UDL N	M&O UDL %	OGR N	OGR %	Difference	PRESENT GRADE DEC 1975 - UOR N	PRESENT GRADE DEC 1975 - UOR %
COL	4,739	7.57	4,276	6.83	-463	4,306	6.53
LTC	10,358	16.55	11,000	17.57	+642	10,153	15.38
MAJ	13,744	21.95	19,204	30.68	+5,460	14,307	21.67
CAPT/LT	33,761	53.93	28,122	44.92	-5,639	37,248	56.42
	62,602		62,602			66,014	

In addition to the needed increase in majors, other conditions were found to exist based on projection data. Examining the projections for the 54 career fields it was noted that some career fields were over graded at various grade levels and that some were undergraded. For example, turning again to the scientific career field, it may be noted in Table 7 that many more majors and lieutenant colonel authorizations are needed for scientific jobs based on the content and responsibility levels of the work performed.

Table 7. AFSC: 26XX TITLE: SCIENTIFIC

Grade	M&O UDL N	M&O UDL %	OGR N	OGR %	Difference	PRESENT GRADE DEC 1975 - UOR N	PRESENT GRADE DEC 1975 - UOR %
COL	12	1.07	6	0.53	-6	6	0.51
LT COL	102	9.09	141	12.57	+39	98	8.28
MAJ	273	24.33	504	44.92	+231	257	21.71
CAPT/LT	735	65.51	471	41.98	-264	823	69.51
	1,122		1,122			1,184	

V. CONCLUSIONS

The officer grade requirements (OGR) project has been one of the most extensive job evaluation research projects in existence. In the course of over 15 years of research and development the technology was used to determine the appropriate grades for 11,825 jobs based on individual ratings in 1963-1966. In the recent development of the technology employing Management Engineering raters, 13,046 individual jobs were rated. Taken together this constitutes over 24,000 jobs which have been evaluated using the OGR technology. When the 1966 projections to 105,908 jobs are coupled with the 1974-1976 projections to 62,602 jobs this results in over 168,000 jobs having been considered by the OGR job evaluation system.

The basic components of the OGR system consist of five job factors (special training and work experience, communication skills, judgment and decision making, planning, and management) judged with benchmark scales, two organization variables, and the supervisor's judgment of the appropriate grade. When entered into a multiple linear regression equation, these variables produce a composite score which can be converted into an appropriate military grade from lieutenant through colonel.

When applied by the METs, this technology will provide a systematic and reliable device for determining officer grade requirements based upon job content and responsibility associated with specific jobs.

Note: This paper has drawn freely from various Air Force Technical Reports regarding the OGR research. A list of key reports is provided in the references.

References

Brokaw, L. D., & Giorgia, M. J. Development of benchmark scales for Air Force officer position evaluation. PRL-TR-66-9, AD-645 055. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division, September 1966.

Christal, R. E. Officer grade requirements project: I. Overview. PRL-TR-65-15, AD-622 806. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division, September 1965.

Christal, R. E. Systematic method for establishing officer grade requirements based upon job demands. AFHRL-TR-75-36, AD-A015 756. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, July 1975.

Hazel, J. T., Finstuen, K., Christal, R. E., & Pope, W. H. Determination of non-aircrew officer grade requirements. AFHRL-TR-77-XX, AD-XXXX XXX. Brooks AFB, Tex.: Occupational and Manpower Research Division (in preparation).

Hazel, J. T. Officer grade requirements project: II. Job descriptions, sample selection, and criterion board. PRL-TR-65-18, AD-629 575. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division, November 1965.

Hazel, J. T., Christal, R. E., & Hoggatt, R. S. Officer grade requirements project: IV. Development and validation of a policy equation to predict criterion board ratings. PRL-TR-66-16, AD-656 125. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division, November, 1966.

Stacy, W. J., Matthews, G. N., & Hazel, J. T. Determination of officer grade requirements by management engineering teams. AFHRL-TR-75-80, AD-A025 309. Lackland AFB, Tex.: Occupational and Manpower Research Division, December 1975.

JOB EVALUATION AS A LINKAGE TECHNOLOGY

MS. LINDA PAPPAS

ALLAN H. FISHER, JR.

FRANK B. MARTIN, JR.

Prepared for:

The Military Testing Association

October 1977

JOB EVALUATION OF A LINKAGE TECHNOLOGY [1]

Objective

This study was a research project conducted for the 1975 Quadrennial Review of Military Compensation (QRMC). The project was designed to test the validity of previously developed job linkages used to compare earnings for "equivalent" military and Civil Service job levels and occupations. The study also provided a demonstration of a method for evaluating job content (difficulty) which could be used both to determine linkages and relative positioning of military and Civil Service occupations. Further, the methodology provides a ready mechanism for a direct comparison of the difficulty of military occupations to the difficulty of jobs in the private sector.

The initial objective of the project was to test the linkages developed in the first QRMC conducted in 1967, as well as other proposed linkages. This analysis included a sample of military (combat arms) jobs for which no counterpart exists in the civilian economy and which were not evaluated by the 1967 QRMC. (The combat arms jobs were not evaluated by the 1967 QRMC due to the limitations of the job evaluation techniques employed.) Analyses were also made of pay grades in the Civil Service to evaluate their job content and relative position with respect to selected military pay grades.

It should be noted that pay comparability between the military and Civil Service and the internal equity of military compensation were not examined in this study. Instead, the focus of this study was on the measurement of job content in military and Civil Service occupations. Comparisons were made between the military service and the Federal Civil Service with respect to the job content within selected pay grades of these two job classification systems.

---

1 This presentation is derived from the Executive Summary of a study conducted for the 1975 Quadrennial Review of Military Compensation by Hay Associates under contract number MDA 903-76-C-0018, An Analysis of Selected Linkages Between Military and Civil Service Occupations. Linda D. Pappas, Allan H. Fisher, Jr., and Frank B. Martin, Jr., April 1976

The task of establishing the relative positioning of pay grades in the military service and the Federal Civil Service entailed the use of the following strategy for research: (a) sampling procedures were used to insure that positions with large numbers of incumbents were represented at each pay grade so that the sample jobs represented the norm of job content at each pay grade; (b) jobs at each level were evaluated using a common (single) standard of measurement, the Hay Method, to permit comparative analyses to be made; and (c) the results of the job evaluation were subjected to both visual inspection and statistical analysis to evaluate linkages and determine the relative positioning of pay grades in the military and Federal Civil Service job classification systems.

Details of the approach are summarized below.

Approach

The approach involved the selection of a representative sample of military and Federal Civil Service positions. A two-stage sampling procedure was employed.

First, certain military pay grades and Civil Service pay grades were selected. A total of seven (7) military pay grades and thirteen (13) Civil Service pay grades were represented. The seven military grades represent a broad spectrum of the military pay structure. The military grades were E-3, E-5, E-7, O-1, O-2, O-5, and O-8. These grades include approximately 41% of the total Armed Forces population. The thirteen Civil Service pay grades were GS-3, GS-5, GS-7, GS-9, GS-14, GS-15, GS-18, WG-5, WG-6, WG-8, WG-10, WS-9 and WS-10. These pay grades represent approximately 46% of all General Schedule (white-collar) employees and 54% of all Wage Grade (blue-collar) employees.

Second, occupations were sampled at each of these military and Civil Service pay grades. For both the military and the Civil Service, jobs were purposely selected to represent occupations with large numbers of incumbents. Hence, jobs were selected to best represent the typical jobs at each pay grade. The military job sample was also selected to represent each of the Armed Services. Further, the military sample included jobs in each DOD occupational category including DOD category "O" (Combat Arms) occupations. Jobs in this category were not evaluated by the 1967 QRMC, because these purely military occupations cannot be evaluated by Civil Service standards. However, the military job sample in the present study excluded

jobs for which the incumbents were in training.

The military and Civil Service job samples were reviewed for representativeness by the QRMC and the U.S. Civil Service Commission, respectively. It was determined that the samples constituted a fair representation of jobs at each pay grade.

A total of 140 military jobs and 193 Civil Service jobs were included in the sample, for a grand total of 333 jobs.[1] Although the number of jobs sampled may appear to be small, the number is quite large compared to other surveys, e.g., the PATC Survey.[2] Further, these jobs were purposely selected to represent occupational specialties with large numbers of incumbents, and hence to provide the basis for estimating the typical job content at each pay grade. In total, the sample positions include 48% of the military population in the seven selected military pay grades, and 42% of the Civil Service population in the thirteen selected Civil Service pay grades. This sample size is adequate for estimating the typical difficulty of jobs at these pay grades in the military and Federal Civil Service.

Written job description materials provided by the Armed Forces and the U.S. Civil Service Commission were used to describe the content of each job.[3] Similar materials were used by the 1967 QRMC in their evaluations. Existing written job description materials were adequate for purposes of this study. The quality of the materials did vary in terms of relevance and completeness. For example, it was found that:

---

1 The term "jobs" is used throughout this report to describe occupations and skill-level combinations.

2 In the 1974 PATC Survey, 84 occupation/skill level combinations comprised the sample used to estimate Federal Civil Service pay requirements for jobs classified under the General Schedule System.

3 The evaluation of O-2 positions was an exception, since job descriptions unique to the O-2 pay grade are generally unavailable.

- The job description information was generally not integrated. For example, military Tables of Organization (T/O's) were separate from the job descriptions;

- The accountability dimension, which has key importance in the Hay evaluation system, was not well specified in many of the job description materials. For example, the military job descriptions did not report the numbers of men supervised or the budget controlled by an officer;

- For some positions, the job description materials appeared to exaggerate the content of the job. For example, some of the Federal Civil Service classification standards listed all duties an incumbent might perform, not the actual duties performed.

- In contrast, the job descriptions for some positions were so brief that it was difficult to understand the full range of job content. For example, the military job descriptions for some officer occupations were only one paragraph long.

However, these problems were overcome in the course of the project. For example, the job description materials were supplemented as needed with additional clarifying information as to the purpose, nature, scope and dimensions of the positions. Military Tables of Organization and information on unit size were used to complement the military job descriptions. Position classification experts in the U.S. Civil Service Commission clarified the problems noted for some Civil Service occupations.

It is important to note that the evaluations of the military occupations in this study may be conservative, in that each military job was evaluated assuming a peacetime environment. This assumption was also made by the 1967 QRMC. As such, combat-related aspects of military jobs were not considered. No weight or score adjustment was made to the military job evaluation scores to reflect these and other conditions unique to military service which increase the difficulty of military occupations.

An Evaluation Committee of experienced private sector consultants was formed to evaluate the 333 positions. These consultants had prior work experience in either the military service or the Federal Civil Service. They also had extensive job evaluation experience in the private sector.

In this application of the Hay Method, job evaluators rated

each military and Civil Service job based on an understanding of the nature of the job. These ratings were made using Guide Charts which represent the three factors of Know-How, Problem Solving and Accountability inherent in each job. These three factors are, in turn, defined by eight dimensions. Know-How is defined by: (1) the extent of knowledge required by the job; (2) breadth of managerial skills; and (3) human relations requirements. Problem Solving is defined by: (1) the degree of original thought required on the job; (2) the degree of limitations imposed on thinking. Accountability is the impact of the job on end results and is defined by: (1) the extent of freedom to act in the job; (2) the degree of primary (vs. shared) accountability in the job; and (3) the magnitude (size) of the job expressed in terms of resources.

In addition to the above factors, consideration was given to Working Conditions when rating blue-collar non-supervisory positions in the military and Civil Service. Working Conditions are defined by three dimensions: (1) the extent of physical effort required by the job; (2) the extent of exposure to hazards; and (3) the quality of the environment, e.g., extent of exposure to noise, fumes, heat, dirt, etc.

The following sequential multi-stage job evaluation process was employed:

- Independent ("raw score") evaluations of each job were made.

- Clarification score evaluations were determined in committee.

- Consensus score evaluations were determined in committee.

- "Sore-Thumbing" consisting of an internal consistency review of consensus scores was done.

- Sponsor review of the evaluations was made. The QRMC staff reviewed evaluations of military postions. Representatives of the U.S. Civil Service Commission reviewed evaluations of the Civil Service jobs; a representative of the Standards Division reviewed evaluations of the General Schedule positions; a representative of the Trades and Labor Section reviewed evaluations of Wage Grade and Wage Supervisory positions.

The evaluation data were compiled, resulting in one final total point score for each of the 333 positions. Summary data were computed, e.g., the median (typical) job difficulty at a particular pay grade.[1] The data were then inspected to determine the relative positions of the military pay grades with respect to the various Civil Service pay grades. A series of statistical tests were also made of proposed linkages between pay grades in the military and Federal Civil Service.

It should be noted that attempts to relate pay grades in the military service to pay grades in the Civil Service are complicated by the fact that different numbers of grades (levels) exist in these two classification systems. A one-to-one mapping of pay grades between the two systems is impossible. For example, the eight officer grades (O-1 through O-8) do not line up with the ten General Schedule grades (GS-7, 9, 11, 12, 13, 14, 15, 16, 17, and 18). Further, the nine enlisted grades do not line up with the fifteen Wage Grades or the nineteen Wage Supervisory grades, or the seven General Schedule pay grades (GS-3 through GS-9). For this reason, attempts to relate the different systems require that estimates be made of the relative position of pay grades in one system compared to pay grades in the other system. Thus, exact linkages should not be expected and relationships like E-3 between WG-5 and WG-6 should be anticipated and readily accepted.

Results

A. Enlisted Level

The present study analyzed the relationships of job content (difficulty) between military enlisted occupations at the E-3, E-5, E-7 levels and Civil Service jobs under both the Civil Service white-collar (General Schedule) system and the blue-collar (Wage Grade/Wage Supervisor) system.

A.1. Relationships Between Pay Grades in the General Schedule and the Enlisted Grades of the Military Service

The ladder in Figure 1 shows the relationship of jobs in

1 The median is a measure of central tendency. Half the jobs are more difficult than the median value; half the jobs are less difficult.

selected levels of the Civil Service General Schedule system and jobs in the military pay grades of E-3, and E-7.

In terms of relative job content, sample jobs in the enlisted grades were evaluated as possessing increased difficulty at each higher level. Sample E-7 jobs were evaluated as more difficult than sample E-5 jobs. Sample E-5 jobs were evaluated as more difficult than sample E-3 jobs. Job content also varied by pay grade for the sample GS jobs. Thus, sample GS-7 jobs were evaluated as more difficult than sample GS-5 jobs, and GS-5 jobs were evaluated as more difficult than GS-3 jobs.

Although the job content in the various pay grades of both systems was evaluated as showing increased difficulty at higher grades, some evidence of overlap in job difficulty between grades was also found in each system. This implies that some jobs have the same content (difficulty), although they are classified in different pay grades. For example, some jobs at the GS-7 and GS-5 levels overlap in the Civil Service system, while some jobs at the E-7 and E-5 levels overlap in the military service.

The following relative positioning of the military enlisted pay grades and the GS pay grades was found by inspection:

- The content of the sample E-3 jobs is similar to the content of the sample GS-3 jobs. The content of the E-5 jobs is similar to the content of the GS-5 jobs. The content of the E-7 jobs is similar to the content of the GS-7 jobs.

- In terms of <u>median</u> job content, the median E-3 job lies near the median GS-3 job; the median E-5 job lies near the median GS-5 job; and the median E-7 job falls near the median GS-7 job. Note that the median job content for each military grade exceeds the median for the corresponding Civil Service grade. Thus, the E-3 median is 9% greater than the median for GS-3; the E-5 median is 7% greater than the median for GS-5; and the E-7 median is 5% greater than the median for GS-7.

- In terms of <u>range</u>, the E-3 and GS-3 jobs have an almost identical range of job content. The relatively narrow range of GS-5 jobs falls within the wider range of E-5 jobs. The range of E-7 jobs falls within the relatively wider range of GS-7 jobs.

MILITARY-CIVIL SERVICE JOB CONTENT COMPARISONS: ENLISTED AND GENERAL SCHEDULE PAY GRADES

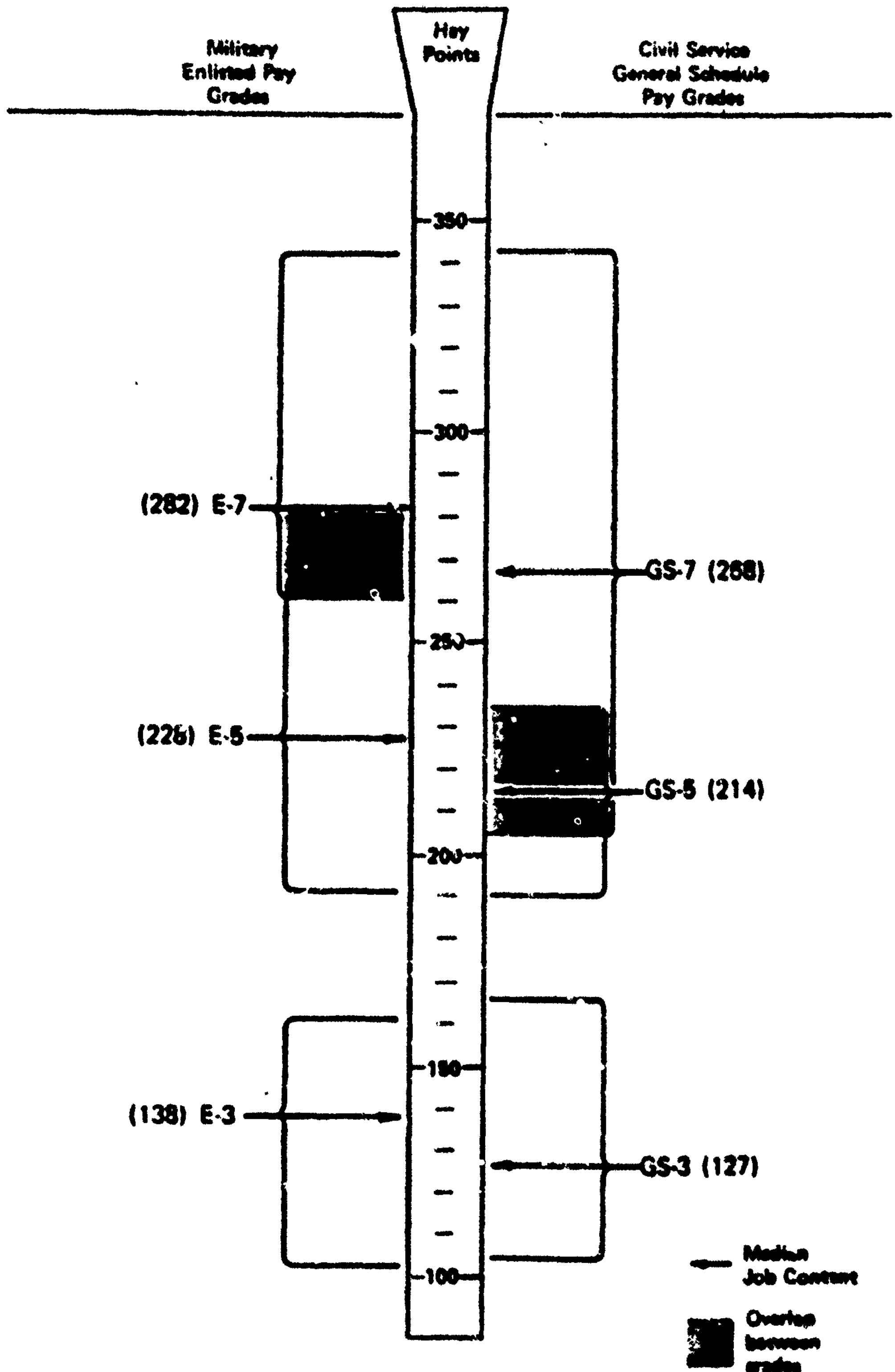


FIGURE 1

Statistical analyses indicated that linkages exist at E-3/GS-3; E-5/GS-5; and E-7/GS-7. The 1967 QRMC had identified a linkage of E-3 to GS-3. The 1967 QRMC also found a linkage of E-5 and GS-5, for the Navy and Marine Corps only. Linkages at these levels were supported by the results of the present study. A statistical linkage of E-7 to GS-7 was also found in the present study.

A.2. Relationships of the Wage Grade System to E-3 and E-5

The ladder in Figure 2 shows the relationship of levels in the Civil Service Wage Grade system and in the military grades of E-3 and E-5.

In terms of relative job content, the E-5 jobs were evaluated as more difficult than the E-3 jobs as noted previously. Similarly, the sample WG-10 jobs were evaluated as more difficult than WG-8 jobs, WG-8 jobs were evaluated as more difficult than WG-6 jobs, and WG-6 jobs were evaluated as more difficult than WG-5 jobs.

Although the job content in the various Wage Grades was evaluated as showing increased difficulty at higher grades, some evidence of overlap in job content between pay grades was noted. Overlap was found between jobs at the WG-5 and WG-6 levels, and between jobs at the WG-8 and WG-10 levels.

Overlap was not found between the E-3 and E-5 levels.

The following relative positioning of military and Wage Grade levels was found by inspection:

- The content of the sample E-3 jobs is similar to the content of the sample WG-5 and WG-6 jobs. In terms of median job content, the median E-3 job lies between the median WG-5 job and median WG-6 job. The E-3 median is 24% greater than the median for WG-5. In terms of range, the range of WG-6 jobs lies totally within the range of E-3 jobs. In contrast, all of the WG-5 jobs were evaluated below the median for the E-3 pay grade.

- The content of the sample E-5 jobs is generally similar to the content of WG-8 and WG-10 jobs. The median E-5 job falls between the median WG-8 job and median WG-10 job. The E-5 median is 22% greater than the median for WG-8. The range of E-5 jobs is similar to the range of WG-10 jobs. In contrast, the range of WG-8 jobs falls at the lower end of the E-5 distribution.

MILITARY-CIVIL SERVICE JOB CONTENT COMPARISONS:
ENLISTED AND WAGE GRADE
PAY GRADES

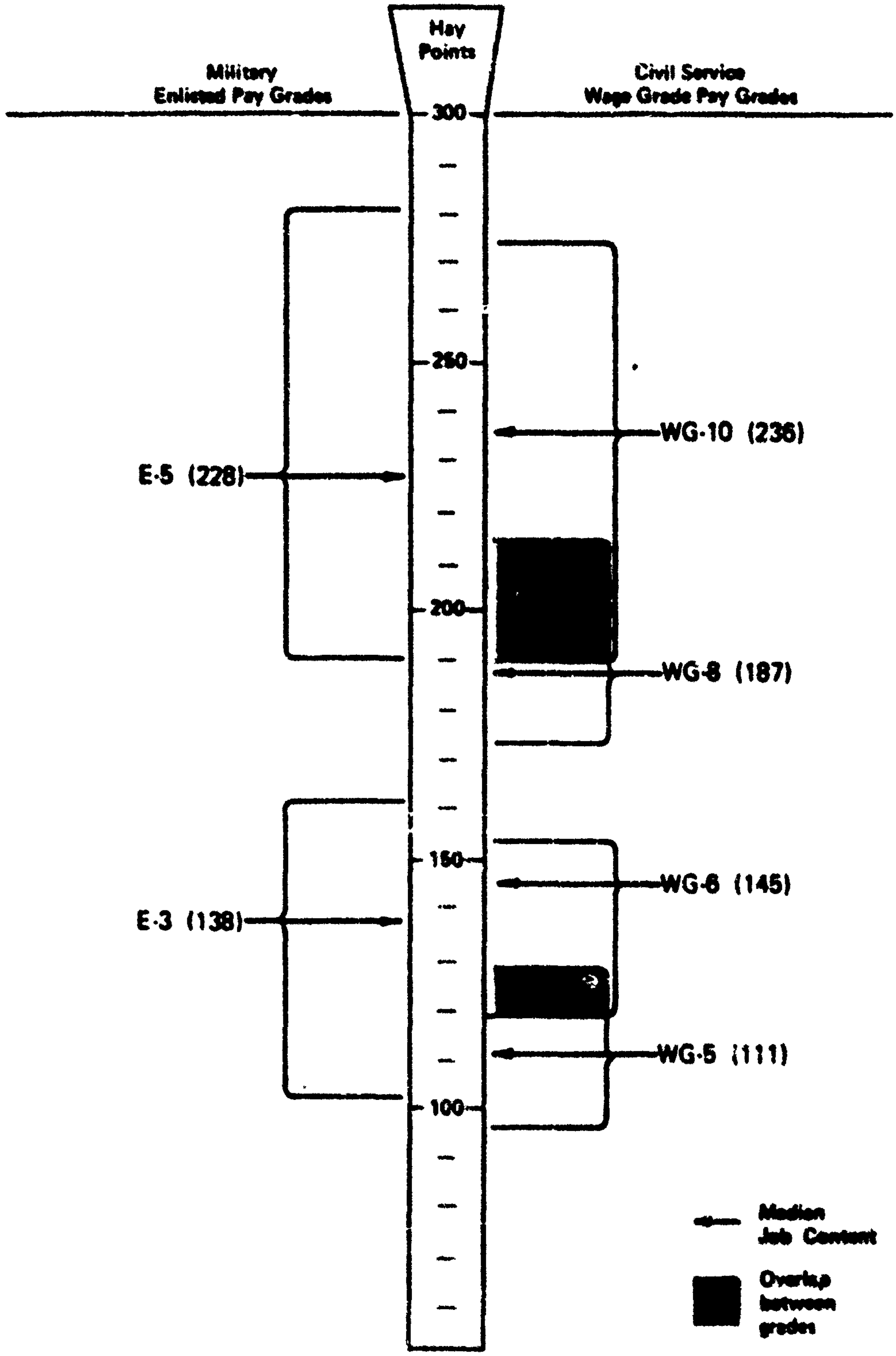


FIGURE 2

Statistical analyses indicated a linkage of E-3 to WG-6. A statistical linkage of E-5 to WG-10 was also found. The 1967 QRMC identified a linkage of E-3 to WB-5 (now WG-5). The 1967 QRMC also found a linkage of E-5 to WB-10 (now WG-10) for the Navy and Marine Corps, but not the Army and Air Force. The E-3/WG-5 linkage found by the 1967 QRMC was not supported by the results of the present study. A linkage was found at the E-5/WG-10 level, although interservice comparisons were not made in this study. As noted above, a link at this level was found only for the Navy and Marine Corps by the 1967 QRMC. The statistical analysis also showed that the E-5 and WG-8 jobs differ in content and that the E-3 and WG-5 jobs differ in content.

A.3. Relationship Between Pay Grades in the Wage Supervisory System and E-7

The ladder in Figure 3 shows the relationship between two levels in the Civil Service Wage Supervisory system and the military grade of E-7.

In terms of relative job content, sample E-7 jobs have previously been shown to be more difficult than sample E-3 and E-5 jobs (See Figure 1 for details). Similarly, the sample WS-10 jobs were evaluated as more difficult than WS-9 jobs.

Although the job content in the two Wage Supervisory pay grades was evaluated as showing increased difficulty in WS-10 compared to WS-9, there was significant overlap.

The following relative positioning of the E-7 pay grade to the two Supervisory pay levels was found by inspection:

- The content of the sample E-7 jobs is similar to the content of the sample WS-9 and WS-10 jobs. In terms of median job content, the median E-7 job lies between the median WS-9 job and median WS-10 job. The E-7 median was 5% greater than the WS-9 median. In terms of range, the range of E 7 jobs lies totally within the range of WS-10 jobs, suggesting an equivalence between these levels. In contrast, 80% of the WS-9 jobs were evaluated below the median for the E-7 pay grade.

Statistical analyses indicated that the sample E-7 jobs were different in difficulty from either the WS-9 or the WS-10 jobs. No statistical linkage was found of E-7 to WS-9, or E-7 to WS-10. These results suggest that the median E-7 job clearly lies within the range of WS-9 to WS-10.

MILITARY-CIVIL SERVICE JOB CONTENT COMPARISONS:
ENLISTED (E-7) AND WAGE SUPERVISORY
PAY GRADES

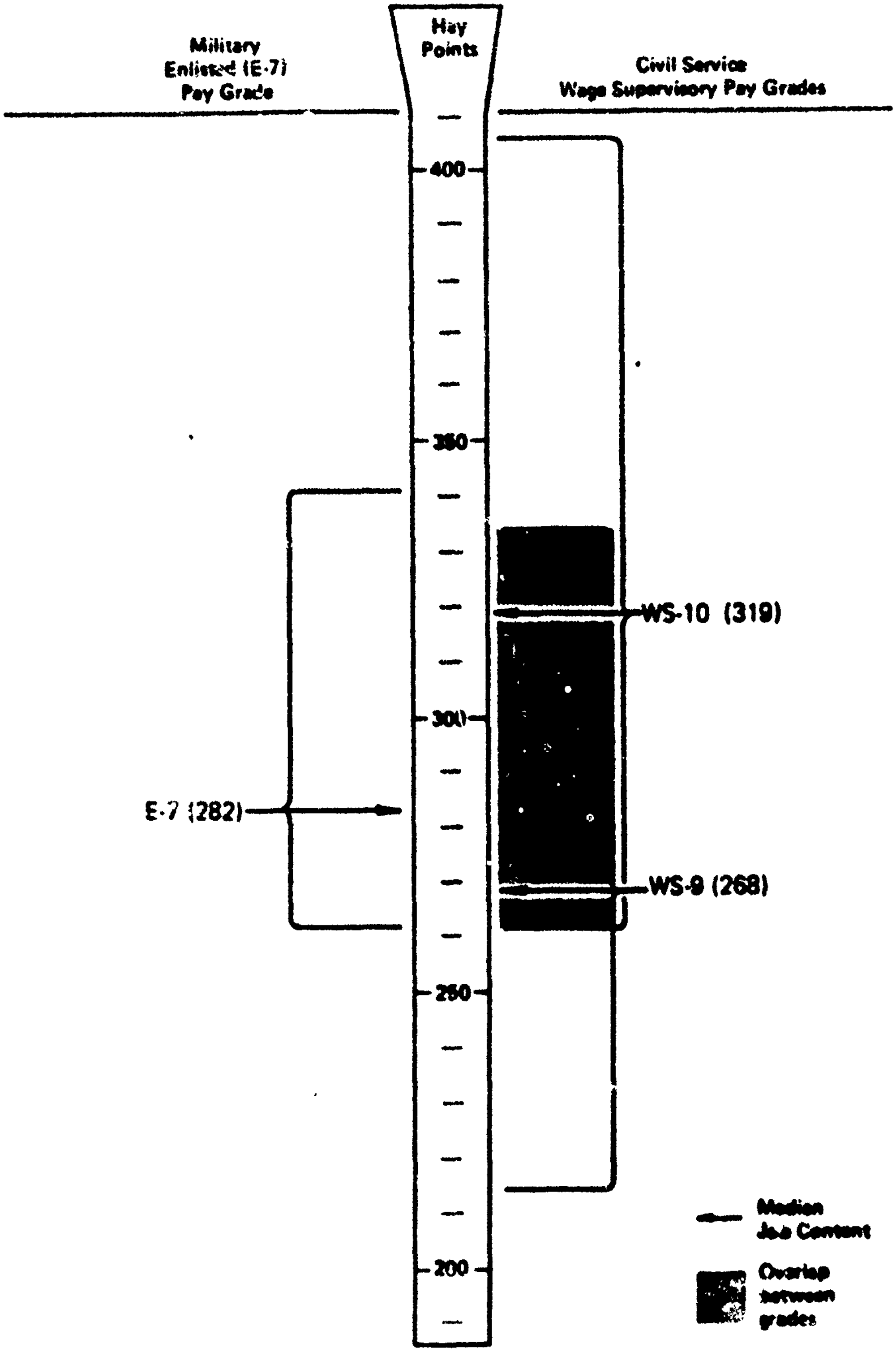


FIGURE 3

B. Officer Level

The present study also evaluated the relationship between pay grades in the General Schedule system and in the military officer grades. Results are discussed separately for the O-1, O-2, and O-5 grades, and for the O-8 pay grade, relative to levels in the General Schedule (GS) pay structure.

B.1. Relationships Between Pay Grades in the General Schedule System and the O-1, O-2 and O-5 Grades

The ladder in Figure 4 shows the relationship of selected levels in the Civil Service General Schedule system and in the military grades of O-1, O-2, and O-5.

In terms of relative job content, sample O-5 jobs were evaluated as much more difficult than sample O-1 and O-2 jobs. The sample O-2 jobs were evaluated as more difficult than the O-1 jobs although the evaluations were quite similar (See statistical test results, below). In the pay grades evaluated under the General Schedule system, the sample GS-15 jobs were evaluated as more difficult than the GS-14 jobs, and the GS-9 jobs were evaluated as more difficult than the GS-7 jobs.

Evidence of overlap in job difficulty was noted in both the military and Civil Service systems. Overlap was found between O-1 and O-2 in the military service, and between GS-7 and GS-9 as well as GS-14 and GS-15 in the Civil Service.

Overlap was not observed between the O-2 and O-5 levels. This is to be expected, given the differences in jobs at the O-5 level compared to the O-2 level.

No jobs in the interim officer ranks were evaluated in this study. It is possible that some overlap exists between positions in adjacent pay grades involving O-2, O-3, O-4, and O-5.

The following relative positioning of the military officer grades and the GS pay grades was found in inspection:

- The content of the sample O-1 jobs is similar to the content of the sample GS-7 and GS-9 jobs. The content of sample O-2 jobs is similar to the content of sample GS-9 jobs. The content of the sample O-5 jobs is similar to the content of the sample GS-14 jobs and some of the GS-15 jobs.

MILITARY-CIVIL SERVICE JOB CONTENT COMPARISONS: COMPANY AND FIELD GRADE OFFICERS

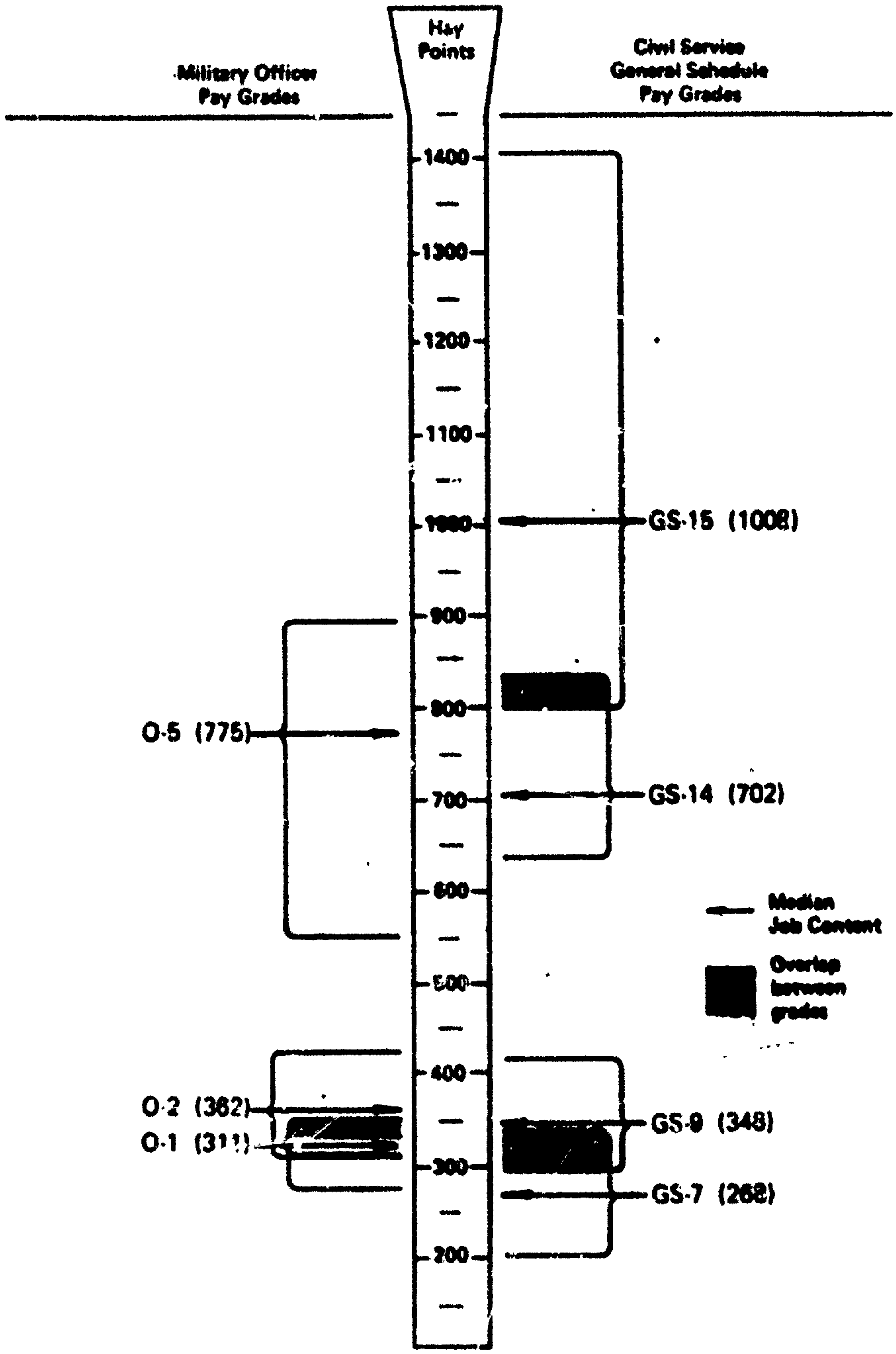


FIGURE 4

- In terms of median job content, the median O-1 job lies between the median GS-7 job and median GS-9 job. The O-1 median is 16% greater than the GS-7 median. The median O-2 job falls above the median GS-9 job. The O-2 median is 4% greater than the GS-9 median. The median O-5 job falls between the median GS-14 job and median GS-15 job. The O-5 median is 10% greater than the GS-14 median.

- With respect to range, O-1 jobs do not fall exclusively into the range of either GS-7 or GS-9; instead O-1 jobs fall in a range bounded by the two ranges defined by the GS-7 and GS-9 job evaluations. In contrast, the range for O-2 and GS-9 are very similar. GS-14 jobs fall totally within the relatively wide range of O-5 jobs.

Statistical analyses indicated linkages of GS-9 to O-1, and GS-9 to O-2. The 1967 QRMC found linkage of O-1 to GS-7. This linkage was not supported by the results of the present study. The statistical analysis indicated that the sample GS-7 positions had less content than the sample O-1 positions.

The finding that GS-9 could be linked to either O-1 or O-2 is explained by the additional finding that the job content of O-1 and O-2 is quite similar. In two of the three tests performed on the data, there was no evidence of a statistically significant difference between O-1 and O-2 evaluations. (As noted earlier, there was considerable overlap in the O-1 and O-2 job evaluations. Their median job content was also similar; the intergrade differential in median job content was only 16%). This finding indicates one problem of attempts to develop exact linkages between classification systems with different grade structures.

Statistical analysis indicated the equivalance of O-5 and GS-14 jobs. A linkage of O-5 to GS-14 was found.

Statistical analysis also indicated that the sample GS-15 positions had more job content than the sample O-5 positions.

3.2. Relationships Between the General Schedule System and the O-8 Pay Grade

The ladder in Figure 5 shows the relationship of selected levels in the Civil Service General Schedule system to the military pay grade of O-8. (The values of O-5, GS-14 and GS-15 from Figure 4 are reproduced in Figure 5 for convenient reference).

MILITARY-CIVIL SERVICE JOB CONTENT COMPARISONS: FLAG RANKS

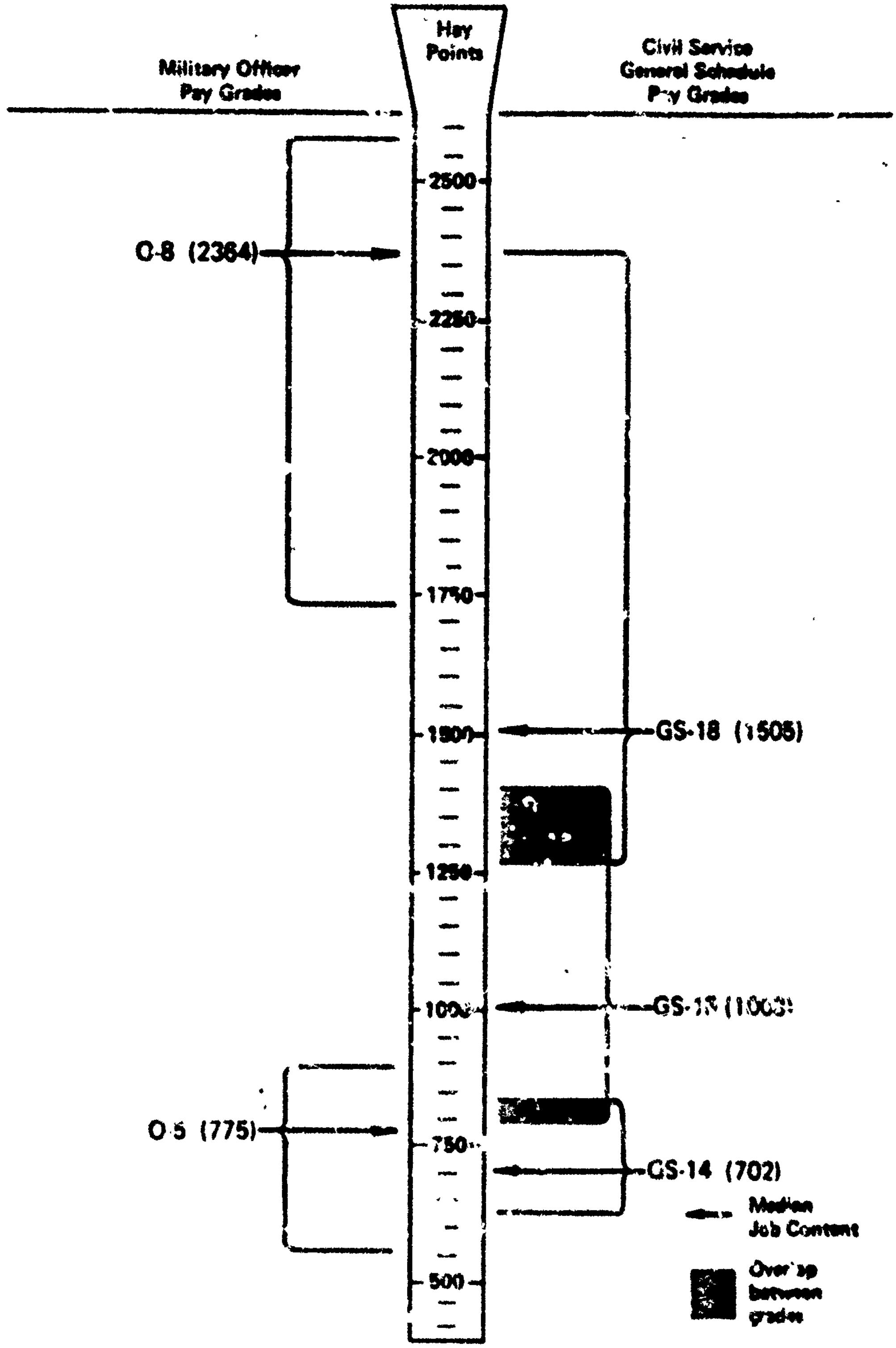


FIGURE 5

The following relative positioning of the military O-8 and the GS-18 pay grade was found by inspection:

- The content of the sample O-8 jobs is similar to the content of the sample GS-18 jobs, to some extent. The more difficult GS-18 jobs are as difficult as some O-8 jobs. However, the less difficult GS-18 jobs were evaluated as far less difficult than any of the O-8 jobs.

- In terms of median job content, the median O-8 job was evaluated as much more difficult than the median GS-18 job. The O-8 median is 57% greater than the GS-18 median.

- In terms of range, there was overlap between the range of O-8 positions and GS-18 positions. The very wide range of difficulty in the sample GS-18 jobs contributed to this finding.

Statistical analyses also indicated that the sample O-8 jobs were different in difficulty from the GS-18 positions. The 1967 QRMC found a linkage of O-8 to GS-18. This linkage was not supported by the results of the present study. The statistical analyses indicated that the sample GS-18 positions have less job content (difficulty), in terms of median value, than do the sample O-8 positions. The results suggest that GS-18 might be a lower-bound or limit with respect to O-8.

Discussion

This study indicated the feasibility of evaluating both military and Federal Civil Service occupations using a point-factor job evaluation approach (the Hay Method). Comparisons of job content between pay grades in the military and Civil Service could be made, since jobs were evaluated on a common standard. In addition, the data were anlayzed to determine overlap in job difficulty within each classification system.

Further, it was possible to analyze "pure" military jobs and compare their difficulty to military jobs which have civilian counterparts. Indeed, a significant finding of this study was the fact that in the enlisted grades and in the company grade officer ranks, the content of "pure" military jobs did not differ systematically from the content of military jobs with civilian counterparts.

Implications of the Findings

In contrast to previous efforts, this study employed a standard statistical criterion to determine if statistical

linkages existed. This approach was more stringent than the criteria applied in previous studies in that the job content in two pay grades had to be quite similar before they could be considered statistically linked.[1]

By applying this statistical criterion of linkage, the present study failed to verify some of the linkages identified by the 1967 QRMC, and used in previous pay comparability analyses. Thus, the present study did not find statistical linkages at these levels:

- O-8 and GS-18
- O-1 and GS-7
- E-3 and WG-5

Statistical linkages of only E-5 and WG-10, E-5 and GS-5, and E-3 and GS-3 were found. Thus, only half of the linkages employed by the 1967 QRMC were supported by the results of the present research. Note that the 1967 QRMC study did not employ this stringent statistical criterion of a linkage. Hence, some differences in the results are not surprising.

SUMMARY OF LINKAGE EVALUATIONS

Source	Proposed Linkage	Statistical Linkages Found
1967 QRMC	O-8/GS-18	NO
	O-1/GS-7	NO
	E-5/GS-5	YES
	E-5/WG-10	YES
	E-3/GS-3	YES
	E-3/WG-5	NO

1 In this study, a linkage required the demonstration that the distribution of job content in two samples (one military grade, one Civil Service grade) was similar. Tests were made to see if the results differed by less than a predefined minimum amount. Only when this condition was satisfied was a statistical linkage said to exist.

The present study also evaluated proposed linkages at levels which were not studied by the 1967 QRMC. The findings support the possibility of a statistical link between O-5 and GS-14, between O-2 (or O-1) and GS-9, and between E-7 and GS-7. A statistical linkage of E-3 and WG-6 was also found.

In total, only half of the linkages hypothesized by the 1967 QRMC were substantiated by the results of this study.

In spite of these findings, the need still remained for the 1975 QRMC to establish the comparability of job content between pay grades in the military and Federal Civil Service classification systems.

Cognizant of the problems associated with attempts to define precise linkages between the military service and the Federal Civil Service, the thrust of this research was redirected toward attempts to determine the relative position of military pay grades with respect to Civil Service pay grades, instead of trying to link military and Civil Service pay grades on a one-for-one basis.

The relative postioning approach was less stringent than the linkage approach. In the linkage approach, the objective was to find two pay grades quite similar in job content. In contrast, the relative positioning approach simply required that Civil Service pay grades be found whose median job content was above or below the median job content of the military pay grade in question. Such Civil Service pay grades would then bracket the military grade and provide the basis for locating the military grade with respect to the two Civil Service grades. Thus, one could position the median job content of the military grade as some fraction of the distance between the median job content of the two Civil Service grades which bracket the military grade.

The relative positioning approach was successfully applied at six military pay grades: E-3; E-5; E-7; O-1; O-2; and O-5.

Results are presented below for the enlisted grades and the General Schedule (white-collar) pay grades:

RELATIONSHIPS OF SELECTED ENLISTED PAY GRADES
TO U.S. CIVIL SERVICE GENERAL SCHEDULE PAY GRADES

Level	Relationship of Median Hay Point Values	Statistical Linkage Findings
E-7	Compared to GS-7 and GS-9, the E-7 median is 18% of the way from GS-7 to GS-9.	E-7 and GS-7 link.
E-5	Compared to GS-5 and GS-7, the E-5 median is 26% of the way from GS-5 to GS-7.	E-5 and GS-5 link.
E-3	Compared to GS-3 and GS-5, the E-3 median is 13% of the way from GS-3 to GS-5.	E-3 and GS-3 link.

The median relationships were such that the median job content for the military grades was bracketed by the median job content for the General Schedule pay grades.

The relationships of the military enlisted grades to Civil Service blue collar pay grades were also determined.

RELATIONSHIPS OF SELECTED ENLISTED PAY GRADES TO U.S. CIVIL SERVICE WAGE GRADE/WAGE SUPERVISORY PAY GRADES

Level	Relationship of Median Hay Point Values	Statistical Findings
E-7	Compared to WS-9 and WS-10, the E-7 median is 24% of the way from WS-9 to WS-10.	WS-10 is an upper bound on E-7; WS-9 is a lower bound.
E-5	Compared to WG-8 and WG-10, the E-5 median is 84% of the way from WG-8 to WG-10.	E-5 and WG-10 are a statistical link. WG-8 is a lower bound on E-5.
E-3	Compared to WG-5 and WG-6, the E-3 median is 79% of the way from WG-5 to WG-6.	E-3 and WG-6 are a statistical link. WG-8 is an upper bound on E-3; WG-5 is a lower bound.

The median job content of E-3, E-5, and E-7 was bracketed by pay grades in the Wage Grade and Wage Supervisory classification systems.

The relationship of military officer grades to pay grades in the Civil Service General Schedule was also examined.

RELATIONSHIPS OF SELECTED OFFICER PAY GRADES TO U.S. CIVIL SERVICE GENERAL SCHEDULE PAY GRADES

Level	Relationship of Median Hay Point Values	Statistical Findings
O-8	The O-8 median exceeds the GS-18 median by 57%.	GS-18 is a lower bound on O-8.
O-5	Compared to GS-14 and GS-15, the O-5 median is 24% of the way from GS-14 to GS-15.	O-5 and GS-14 are a statistical link; GS-15 is an upper bound on O-5.
O-2	Compared to GS-9 and GS-11*, the O-2 median is 14% of the way from GS-9 to GS-11.	O-2 and GS-9 are a statistical link.
O-1	Compared to GS-7 and GS-9, the O-1 median is 54% of the way from GS-7 to GS-9.	GS-7 is a lower bound on O-1. GS-9 and O-1 are a statistical link.

* The GS-11 median value was taken from research conducted in another study.[1]

1 No other GS-11 estimate exists for comparison purposes. However, caution is recommended in the use of this value because the sample size is small (N-5). Further, different criteria were established in each study for the representativeness of the PATCO categories, and different population frames were used.

It was possible to bracket the military grades of O-1, O-2 and O-5 using the General Schedule pay grades noted in the table.

These relationships may be used for estimating grade comparability between the military service and the Federal Civil Service.

Summary

In summary, a limited number of whole grade linkages are identifiable between military and Civil Service pay grades. However, for many military and Civil Service grades the identification of whole grade linkages is not possible. Therefore, the relative position of pay grades in the two systems is important to consider.

The application of the point-factor job evaluation methodology employed in this research permits the identification of both the linkage of whole military and Civil Service grades and the relative positioning of military and Civil Service grades.

Finally, even though it was not a part of the research effort, it should be noted that this job evaluation approach permits the linkage of military jobs to private sector jobs. Virtually hundreds of private sector firms employ this job evaluation approach and an extensive private sector job evaluation and salary data base exists.

WHITE-COLLAR JOB EVALUATION AND PAY SYSTEMS

By: Rosemary Storm

CURRENT INITIATIVES

In the Federal civilian white-collar world, job evaluation and pay are inexorably linked. This paper will set forth the existing statutory framework, some relatively recent history on pay comparability, and current initiatives for improvement.

JOB EVALUATION

Job evaluation in the Federal Government is probably the most sophisticated, detailed and studied in the world. A wido range of occupations is required to accomplish varied Federal activities. Under law, there must be equal pay for substantially equal work across the many agency lines as well as across occupational lines.

Over 1.3 million white-collar positions are classified under the 18-grade General Schedule. The 18 grade levels and their definitions are set in law. Authority to classify positions is vested in the head of each agency, and positions must be classified in accordance with published standards issued by the Civil Service Commission.

In 1977, Federal agencies began implementation of the new Factor Evaluation System--an improved, standardized, factor/point methodology for classifying non-supervisory positions in grades GS-1 through GS-15--after several years of CSC development, testing, modification and retesting. Key elements of the new Factor Evaluation System include a common set of nine factors, defined degrees of each factor, benchmark descriptions of representative positions at various grades, and a conversion chart that translates total points into GS grades.

The overall reaction to the Factor Evaluation System is clearly favorable. It is relatively simple to apply and is easily understood by classifiers, managers, and employees. Most commentators agree that it provides accurate grades, improved alignment across occupational and agency lines, and better documentation of classification decisions.

An accelerated standards production program is underway, using the new methodology. In FY 1977, the Civil Service Commission issued initial standards covering 52,000 positions. Plans are to complete a basic set of new standards within five years.

PAY

The Federal pay comparability process was developed in the late 1950's and early 1960's. It involved four years of data gathering by the Bureau of Labor Statistics (1958-1962), culminating in President Kennedy's urging salary reform and the comparability principle, and the Congress' enacting the Federal Salary Reform Act of 1962. At the time of enactment, Federal pay was so far below comparability that comparisons could be made in rather gross terms.

In 1969 full comparability was reached, and more precise measurement became crucial. There were also changes in the national labor market and in the Federal work force that called into question some aspects of white-collar pay comparisons.

In 1973, a General Accounting Office report pointed out the following needs for improvements: (1) more emphasis on pay research; (2) broader coverage of the PATC survey (this is the Bureau of Labor Statistics survey of professional, administrative, technical, and clerical jobs which is used as a data base to set Federal white-collar pay); and (3) broader industrial scope of the PATC survey. Then, too, other criticisms began to mount from both inside and outside Government. These criticisms generally said that pay for some jobs was too high, and pay for some other jobs was too low. In 1974, the Civil Service Commission undertook research projects in two major categories; (1) improving the present system, and (2) exploration of other methods to get to closer pay comparability.

It was against this background that President Ford, in his FY 1976 budget, announced plans to establish a blue ribbon panel to make policy recommendations to him on Federal pay. Vice President Rockefeller was appointed to head this panel.

The overall theme of the recommendations was support for the comparability principle to set pay. The specific recommendations were of two kinds: Those that could be implemented without legislation (e.g., improved statistical techniques) and those that would require legislation. The major recommendations that would require legislation include:

- splitting the current monolithic General Schedule into two basic Services: a Clerical and Technical Service with local pay schedules and a Professional and Administrative Service with national pay schedules;

- authorizing special occupational services when the regular service hampers management's ability to recruit, retain, or manage a well-qualified work force;

- authorizing use of State and local government pay data in Federal pay surveys;

- conducting major pay surveys less frequently than annually (but using a statistical indicator to adjust pay in intervening years);

- combining or eliminating separate Federal civilian pay systems as needed;

- studying and developing a pay advancement system for professional and administrative employees based on quality of performance; and

- developing and testing methodologies for extending the principle of comparability to benefits as well as pay (total compensation comparability).

Methodologies for achieving a total compensation comparability system are now being tested. The Bureau of Labor Statistics has conducted successful preliminary tests of benefits data collection. Expanded testing is planned for FY 78. This is a project of tremendous technical difficulty. It has stirred considerable interest inside and outside Government, and we look forward to viewing the results of the further testing and developmental work.

On May 27, 1977, President Carter established a Federal Personnel Management Project to make a top-to-bottom study of Federal personnel management as a part of his reorganization effort. A report to the President is due in November 1977. A legislative proposal to accomplish those recommendations submitted by the 1975 President's Panel on Federal Compensation was sent to the Personnel Management Project for review and consideration. As a part of this project, just four short days ago, an options paper on job evaluation, pay and benefits was issued to the public. Eight issues are discussed, with various options for each issue. The issues are:

1. Should the Government extend its pay comparability policy to include benefits as well as pay (total compensation comparability)?
2. What methods of measurement should be used in comparing Federal benefits with non-Federal benefits?
3. Should there be central authority for granting benefit changes to Federal employees? If so, where should it reside?
4. To bring about closer pay comparability with other employers,

(a) Should the General Schedule be divided into two or more homogeneous occupational groupings with separate classification and pay systems?
(b) Should some or all of the General Schedule work force be paid on the basis of local rates?
(c) Should the President be authorized to establish and abolish special pay systems for specific occupations or groups of occupations?

5. Should comparisons include State and local government employees for purposes of establishing comparability?
6. Can the principle of merit pay be used to improve and reward employee performance?
7. What improvements are needed in the job evaluation process?
8. What should be done about the changing relationship between blue-collar and white-collar pay rates?

Views from interested parties will be considered and recommendations formulated for transmittal to the President in November--and for possible inclusion in his budget message in January 1978.

CERTIFICATION & LICENSURE PROGRAMS FOR OCCUPATIONAL SKILL DOCUMENTATION

Roger G. Goldberg

The Role of the Defense Activity for Non-Traditional Education Support (DANTES) in the Provision of Skill Documentation Programs

The Defense Activity for Non-Traditional Education Support (DANTES) was established by Congress in June 1974 to provide independent study and examination programs for military personnel.

Through memoranda of understanding executed with approximately 65 regionally accredited colleges and universities, we publish a catalog and provide descriptive information on over 10,000 independent study courses from the high school and undergraduate level through graduate study and professional continuing education. The availability of this variety of educational programs is designed so that military personnel may be able to continue to pursue their educational objectives no matter how isolated or remote the duty station and inaccessible resident study.

Through contractual agreements with organizations such as the Educational Testing Service, College Entrance Examination Board, American Council on Education and American College Testing we provide examination programs such as the GED High School Equivalency Examinations, the College Level Examination Program (CLEP), the DANTES Subject Standardized Tests (DSST's), Graduate Record Examinations and many others. Examinations that provide credit for high school course equivalency, high school equivalency certificates, college entrance and credit and graduate admissions.

Through approximately 880 testing sites located throughout the world over 275,000 examinations are administered yearly.

In the college level credit examination programs, the CLEP and DSST's, 191,000 examinations were administered through August of this year. Approximately 101,000 individuals passed the examinations with scores, at or above the level, established by the American Council on Education for the awarding of academic credit. These examinations provided a potential of 302,700 credit hours at a cost of $2.87 per credit hour. This compares with the civilian administered examination program cost of approximately $7.00 per credit hour or the inservice tuition assistance cost of approximately $25.00 per credit hour. We estimate, based on these two examination programs alone, a potential cost savings/cost avoidance to DOD of over $5,750,000.

In addition to the academically oriented examination programs, DANTES has been expanding both the role and number of vocational-technical and para-professional examination programs within DOD. We view our activities in this area as supportive of the various service sponsored job recognition/skill documentation programs and anticipate significant growth in this program area.

Capitalizing on our unique role as the provider of examination programs for the Department of Defense, we have developed memoranda of understanding with a number of nationally recognized certification organizations

allowing for DANTES Test Control Officers, world-wide, to administer technical and professional certification examinations.

To date, the following organizations have signed memoranda with DANTES and their certification by examination programs are being administered by DANTES Test Control Officers at military installations.

American Association of Medical Assistants

Certified Medical Assistant

Basic
Administrative
Clinical

American Medical Technologists

Medical Technologist
Medical Laboratory Technicians
Registered Medical Assistant

Institute for the Certification of Engineering Technicians

Certified Associate Engineering Technician
Certified Engineering Technician
Certified Senior Engineering Technician

Architectural & Building Construction Technology
Civil Engineering Technology
Electronics Engineering Technology
Fluid Power Engineering Technology
Industrial Engineering Technology
Mechanical Engineering Technology
Metallurgical Engineering Technology
Geotechnical Engineering Technology
Constructions Materials Testing
Electrical Power (Production/Transmission/Sub-Station Distribution)

National Institute for Automotive Service Excellence

Certified General Automobile Mechanic

Engine Repair
Automatic Transmission
Manual Transmission and Rear Axle
Front End
Brakes
Electrical Systems
Heating and Air Conditioning
Engine Tune-Up

Certified Heavy-Duty Truck Mechanic

Gasoline Engines
Diesel Engines
Drive Train
Brakes
Suspension and Steering
Electrical Systems

Certified Body Repairer
Certified Painter & Refinisher

Institute for Certification of Computer Professionals

Certified Data Processor
Certified Computer Programmer

Business Programming
Scientific Programming
Systems Programming

International Society for Clinical Laboratory Technology

Registered Medical Technologist
Registered Laboratory Technician

National Registry of Emergency Medical Technicians

Registered Emergency Medical Technician

Ambulance
Non-Ambulance
Paramedic

Institute of Certified Professional Managers

Certified Manager

In addition to the above organizations, agreements have been effected or are being negotiated with the American Society for Quality Control, the American Registry of Radiologic Technologists, the Academy of Certified Social Workers and with approximately thirty-five states using a uniform examination for the licensing of real estate salespersons and brokers.

The preponderance of state licensing examination programs are highly state-specific, in terms of the various requirements for licensing, the design of the test instrument and in the jurisdictional acceptance of the examination. The lack of formalized reciprocity agreements and/or comity amongst the states in regard to state licensing examinations has

limited the utility of licensing examination programs for DANTES administration. The relatively uniform examination procedures employed in the field of real estate licensing adapt well to our testing program and hopefully additional uniform licensing examinations will be identified and incorporated into our testing program in the future.

In the meantime, however, we are attempting to provide the Armed Forces Education Centers with information on state licensing programs. Recently, Florida State University, under a DANTES contract, completed a survey of the agencies within every state that had proponency for the regulation and licensing of occupations. The results of the survey will be published as the Directory of Selected Licensable Occupations and will be distributed to all Armed Forces Education Centers world-wide. Over thirty occupations will be represented in the Directory and information provided on age, experiential, education and residency requirements, examination and licensing fees, dates and sites of examination administrations, retest policy and procedures, reciprocity and comity agreements and contact points for further information.

The Directory represents a first time effort to gather this type of information on a systematic and nationwide basis. In the early research phase of the project the Department of Labor was contacted and informed of our intentions and has since expressed keen interest in providing funding for this effort on a continuing basis.

This past year DANTES conducted a pilot project regarding the use of the National Occupational Competency Testing Institute (NOCTI) Examinations. Originally developed under an HEW contract, the NOCTI Examinations are designed to certify as vocational-technical instructors those individuals with many years of on-the-job experience, but little formal education.

The NOCTI Examinations are available in a number of vocational-technical skill areas and consist of a written multiple-choice examination and a practical demonstration of skills examination supervised and evaluated by a certified NOCTI evaluator.

The NOCTI Examinations were administered to 159 military examinees across the country and the scores provided to the Los Angeles City Colleges and the City Colleges of Chicago for credit evaluation. Credit recommendations from the City Colleges of Chicago ranged from 0-38 semester credit hours with the mean recommendation of 15 hours.

During this fiscal year DANTES will be exploring with the Department of Labor and the individual services, the adaptability of the NOCTI Examinations for credit towards the formalized educational requirements of the service sponsored apprenticeship programs.

Each apprenticeship program registered with the Department of Labor usually has a formal training requirement of 144 hours of instruction per

apprenticeship year. The apprenticeship programs registered by the military to date, have achieved recognition by the Department of Labor, of service school technical training as fully meeting program requirements. However, this aspect of the apprenticeship programs would exclude individuals who have not attended or completed a service school and who gained their skill proficiencies through on-the-job training and experience. We are hopeful that the Department of Labor will allow for the use of the NOCTI Examinations in lieu of formalized training so that all qualified military personnel may participate in the various apprenticeship programs.

Acceptance by the Department of Labor of the NOCTI Examinations would provide the NOCTI Exams with a potential dual capability; apprenticeship credit and academic credit. This dual capability would further expand career enhancement opportunities for service personnel as well as complement existing programs for obtaining academic credit. Military personnel may already earn academic credit for their MOS (Military Occupational Skill) or NEC (Naval Enlisted Code), as well as for the courses taken in service schools. With the combination of programs currently available, military personnel often find that after several years of service an Associate degree or Baccalaureate degree is within reach.

If the Office on Educational Credit of the American Council on Education can successfully establish academic credit recommendations for apprenticeship programs and certification examinations, as they are now investigating, military personnel will benefit from increased opportunities for academic achievement, peer recognition and for greater meaningful alternatives for personal growth and career enhancement.

submitted by: Roger G. Goldberg
Head, Certification & Licensure Programs Branch
Defense Activity for Non-Traditional Education Support
Ellyson Center
Pensacola, Florida 32509

DANTES Discussion Panel

NATIONAL APPRENTICESHIP PROGRAM

Program Manager: Ken Tellier

CONCEPT/ PURPOSE To provide a brief description of the National Apprenticeship Program for active-duty personnel.

BACKGROUND On 24 March 1976, the Secretary of Labor and the Secretary of the Navy signed an agreement which would permit active-duty Navy personnel to complete apprenticeships in civilian trades. In accordance with this agreement, the Bureau of Apprenticeship and Training of the Department of Labor recognizes certain Navy skills as civilian "apprenticeable occupations." Navy persons who achieve documented levels of experience and training in these skills are recognized by the Department of Labor. The Navy, on the other hand, agrees to civilian standards of training and experience and administers the program in a manner acceptable to civilian industry.

STATUS/ DISCUSSION CNET has received approval from the Department of Labor to register and administer apprenticeships for five Navy skills:

Office Machines Mechanic
Watch Clock Repairer
Commercial Photographer
Camera Repairer
Hotel and Restaurant Cook

As of 15 September 1977, 123 personnel in the IM, PH, and MS ratings have been registered in the five trades listed above with the Department of Labor. To qualify for registration, the Navy person must have completed from 288 to 432 hours of formal instruction in his trade (usually completed at an applicable Class A or Class C school). To qualify for the completion certificate awarded by the Department of Labor, the registered apprentice must complete from 4000 to 6000 hours of documented work experience in the trade.

There is little likelihood that a Navy person can complete an apprenticeship within a four-year enlistment. Progress in an apprenticeship goes hand-in-hand with advancement in rating. After a Navy person has completed the required work experience and is regarded as a journeyman by the Department of Labor, he not only has civilian proof of competence in a trade, but he is a more competent Navy person.

CHAPTER II

NATIONAL APPRENTICESHIP PROGRAM

EXHIBITS

CHAPTER II

NATIONAL APPRENTICESHIP PROGRAM

201. INTRODUCTION

1. The National Apprenticeship Program is a component program of the Navy Campus For Achievement established by written agreement between the Secretary of Labor and the Secretary of the Navy on 24 March 1976.

2. The objectives of a National Apprenticeship Program are to develop highly skilled Navy-oriented journeymen who will continue to utilize their technical skills and knowledge within the Navy, and who will incidentally qualify for employment in a recognized civilian trade after expiration of enlistment or upon retirement. Adherence to the standards of an apprenticeship program will also reinforce efforts leading to advancement in rating by the individual apprentice. The success of the program will gain wide recognition of the worth of Navy training and experience.

3. The Chief of Naval Education and Training will identify the trades to be introduced as apprenticeable occupations within the active duty Navy. The identification of a trade for an apprenticeship will depend upon the following: the conditions and trends of the national labor scene; the assurance that selected and registered active duty apprentices will receive work experience and related instruction similar to that received in the civilian sector; the availability of facilities and supervisory personnel for on-the-job training and related instruction; and the assurance that administrative procedures and controls will be exercised in a manner that will earn the confidence and respect of the civilian trade sector.

202. TERMINOLOGY. To be credible and usable, the National Apprenticeship Program for active duty personnel is obliged to utilize the terminology employed by the Bureau of Apprenticeship and Training of the U.S. Department of Labor. The following terms are the same as used in civilian apprenticeship programs:

1. Apprentice. A person who is properly registered with the Bureau of Apprenticeship and Training of the Department of Labor. It is to be noted that the words "apprentice" and "apprenticeship" in civilian programs are not the same as the apprenticeship associated with pay grade E-2.

2. Journeyman. A person who has satisfactorily completed an apprenticeship and who has been awarded a certificate attesting to this completion. Navy personnel completing an authorized apprenticeship

under the National Apprenticeship Program will be awarded a Certificate of Apprenticeship Completion by the Bureau of Apprenticeship and Training, Department of Labor.

3. Registration. The action by which a qualified individual is earmarked by the Bureau of Apprenticeship and Training of the Department of Labor as an apprentice. Registration of an individual can only be accomplished by the commanding officer, officer in charge, or director of the Class "A," Class "C," or Class "J" school providing related instruction for the apprenticeable trade. Registration is accomplished by completing an Apprentice Registration Application, CNET Form 1560/1, and by issuing a work experience log to the qualified registrant.

4. Work Experience. Verified on-the-job participation in the skills required by an apprenticeable trade. It is the responsibility of the individual apprentice to record the hours of work experience on a Work Experience Hourly Record, CNET Form 1560/3, and have such entries verified weekly by the signature of the leading petty officer or work center supervisor. Work experience is not to be confused with related instruction.

5. Previous Work Experience. Verified on-the-job participation in the skills required by an apprenticeable trade which were completed prior to registration. Hours of previous work experience, if any, are entered in the apprentice registration application only by the commanding officer, officer in charge, or director of the Class "A," "C," or "J" school at the time of registration.

6. Related Instruction. The formal and/or classroom training acquired at an applicable Class "A," "C," or "J" school, which must be completed as a condition of registration, and which provides the potential apprentice with the required background knowledge and information of the trade Related instruction is not interchangeable with work experience.

7. Work Experience Log. A booklet issued to an apprentice at the time of registration and held thereafter as a personal possession. The work experience log shall be reviewed twice yearly during apprentice progress interviews by the most accessible Navy Campus education specialist, with particular emphasis upon the Work Experience Hourly Record, CNET Form 1560/3. The work experience log identities the apprenticeable trade and contains the following parts/documents:

a. Information for the apprentice.

b. National Apprenticeship Standards for the U.S. Navy.

c. Work Processes Schedule for the trade.

d. Apprentice Progress/Status Report (CNET Form 1560/2).

e. Work Experience Hourly Record (CNET Form 1560/3).

f. Apprentice Registration Application (CNET Form 1560/1). The original of this form is inserted at time of registration.

g. Other documentation. (For example, copies of the service record, page 4; letters from previous employers attesting to previous work experience; copies of documents from non Navy schools attesting to qualifying related instruction).

8. Work Processes Schedule. A listing of the skill areas within an apprenticeable trade, together with the hours of work experience assigned to each skill area (often referred to as a "work experience plan"). The work processes schedule, which is contained in the work experience log, tells the apprentice how many hours of experience must be completed in each skill area of the trade. It is a breakdown of the work experience to be completed during the term of the apprenticeship. Before an entry is made in the work experience hourly record, the apprentice must refer to the work processes schedule in order to identify the skill area in which the work experience has been completed.

9. Work Experience Hourly Record. A form contained within the work experience log, which is used for the entry and verification of completed hours of on-the-job skills.

10. Certificate of Apprenticeship Completion. A document issued by the Bureau of Apprenticeship and Training of the Department of Labor attesting to the fact that an individual has completed the apprenticeship.

203. REGISTRATION PROCEDURES

1. In order to qualify for registration, an individual must have graduated from the Class "A," "C," or "J" school applicable to the apprenticeable trade and must be serving, or about to serve, in an authorized apprenticeable trade. Authorized apprenticeable trades are listed in paragraph 208.

2. The commanding officer, officer in charge, or director of an applicable Class "A," "C," or "J" school is authorized to waive the Navy school requirement if the applicant can provide documentation of satisfactory completion of the required hours of related instruction at an Army, Air Force, Marine Corps, or civilian school. The Department of

Labor requires 144 hours of related instruction for each 2000 hours of an apprenticeship. Therefore, a 6000-hour apprenticeship will require 432 hours of documented related instruction; an 8000-hour apprenticeship, 532 hours, and a 4000-hour apprenticeship, 388 hours.

3. The registration process begins by submitting an Apprentice Registration Application, CNET Form 1560/1 (Exhibit II-A), to the commanding officer, officer in charge, or director of the Class "A," "C," or "J" school providing the required related instruction. Registration is usually accomplished in person; however, registration may also be accomplished by mail if the eligible applicant sends an original and two copies of the application, with the "Applicant Information" entered, to the applicable Class "A," "C," or "J" school.

4. Personnel who register by mail and who are also eligible for previous work experience credit in the trade should include the following documents: (See Exhibit II-B for a recommended forwarding letter)

a. A reproduced copy of page 4 of the service record which displays assignment to a Navy Enlisted Classification (NEC) translatable to previous work experience. See paragraph 208 for NEC's which are authorized for translation into credit for previous work experience.

b. Original letters or similar documentation from previous employers which attest work experience in a skill or skills associated with the apprenticeable trade.

5. The commanding officer, officer in charge, or director of the appropriate school completes the bottom half of the apprentice registration application. An applicant may be credited with 1000 hours of previous work experience for every full year that the applicant's service record (page 4) indicates assignment to an NEC cited in paragraph 208 for the apprenticeable trade. 1000 hours of previous work experience may also be credited for every full year of work experience in the trade which is verified by an original letter from a previous employer.

6. Credit for previous work experience cannot exceed more than 50 percent of the term of the apprenticeship, i.e., no more than 3000 hours of previous work experience can be credited to a 6000-hour apprenticeship. Portions or fractions of years of previous work experience will not be credited. Only full years will be translated and credited.

7. Upon completion of all items on the apprentice registration application, the commanding officer, officer in charge, or director of the school will issue/mail the appropriate work experience log to the

applicant. The following documents will be inserted in the log before issuance/mailing:

a. One copy of the completed apprentice registration application.

b. Service record pages and letters, if any, cited in 4 above.

8. Two copies of the completed apprentice registration application will be forwarded to the Chief of Naval Education and Training.

204. CANCELLATION OF REGISTRATION

1. A registration will be canceled for any one of the following reasons:

a. Request of the apprentice.

b. A rating in the lower 50 percent in "Professional Competence" on the Enlisted Performance Evaluation.

c. Discharge or release to inactive duty.

d. Termination of work experience in the apprenticeable trade for a period of more than one year.

e. Failure to report to a Navy Campus education specialist for twice a year progress interviews unless the apprentice has requested suspension of registration.

2. Cancellation of registration is accomplished through submission of an appropriately checked apprentice progress/status report (Exhibit II-C) signed by one of the following personnel:

a. The commanding officer or officer in charge of the apprentice.

b. A Navy Campus education specialist.

c. The Chief of Naval Education and Training.

d. The apprentice.

3. Cancellation of registration is tantamount to removal from the apprenticeship program. Once a registration is canceled, an individual can reenter the apprenticeship program only by reapplying for registration. The Chief of Naval Education and Training will adjudicate all cases of application for reregistration and will determine whether reregistration of an active duty member will be permitted.

4. Cancellation of registration should not be confused with suspension of registration.

205. SUSPENSION OF REGISTRATION

1. Suspension of registration is temporary. Suspension retains the apprentice in a temporary inactive status for no more than one year but still enables the apprentice to accumulate hours of work experience for entry and verification on the work experience hourly record.

2. Suspension is accomplished by submission of an appropriately checked apprentice progress/status report (Exhibit II-C) to the Chief of Naval Education and Training signed by one of the following personnel:

a. The commanding officer or officer in charge of the apprentice.

b. A Navy Campus education specialist.

c. The Chief of Naval Education and Training.

d. The apprentice.

3. Suspension will be granted by the Chief of Naval Education and Training and the Department of Labor for any one of the following reasons:

a. If the apprentice is unable to complete, for reasons beyond control, work experience in the apprenticeable trade for a period of one year or less; that is, hospitalization, orders to light duty, assignment to duties not related to the trade in which registered.

b. If the apprentice is unable to report for a required apprentice progress interview because of operational requirements or because a Navy Campus education specialist is not available.

4. A suspension will be lifted if the apprentice resumes work experience in the apprenticeable trade within one year after date of

suspension and reports to a Navy Campus education specialist. The Navy Campus education specialist will submit an appropriately checked and dated apprentice progress/status report and will include hours of work experience completed during the suspension, if any.

5. Suspension does not require re-registration. The Chief of Naval Education and Training will examine all cases of repetitious suspensions and will determine whether cancellation is justified.

6. An apprentice is urged to earn work experience hours, however small, during a suspension.

206. RESPONSIBILITIES FOR THE APPRENTICESHIP PROGRAM

1. The Chief of Naval Education and Training

a. Function as the single point of contact with the Department of Labor regarding all policy and management aspects of the Navy's National Apprenticeship Program for active duty personnel.

b. Provide policy guidance for the operation and management of the National Apprenticeship Program for active duty personnel.

c. Identify the trades which are to be introduced as apprenticeable trades within the active duty Navy.

d. After consultation with the Department of Labor, assign responsibility for promulgation of "work processes schedules" for designated apprenticeable trades and ensure the introduction of these schedules into appropriate records.

e. Evaluate the overall effectiveness of the Navy's National Apprenticeship Program for active duty personnel and ensure that acceptable levels of proficiency for apprenticeable trades are being met.

f. Screen and forward apprenticeship registration, suspension, cancellation, and completion actions and individual apprenticeship progress reports to the Department of Labor as required.

2. The Chief of Naval Education and Training Support

a. Arrange for the printing, stocking, and distribution of the forms prescribed for the National Apprenticehsip Program for active duty personnel.

b. Ensure the indoctrination and effectiveness of Navy Campus education specialists regarding the procedures, controls, and action required to provide efficient field management and maximum assistance to commanding officers regarding the National Apprenticeship Program for active duty personnel.

3. Commanding Officers, Officers in Charge, or Directors of Class "A," "C," or "J" Schools

a. Register volunteer members as apprentices if they have completed the required related instruction. The following will be regarded as required related instruction:

(1) Successful completion of the course appropriate to the trade listed in paragraph 208; or

(2) Satisfactorily documented completion of the required hours of related instruction at an Army, Air Force, Marine Corps, or civilian school. See paragraph 203.2 for the hours of required related instruction earned at a non Navy school.

b. Ensure that eligible registrants are credited only with documented hours of previous work experience in accordance with paragraphs 203.4 through 203.6.

c. Issue the appropriate work experience log to volunteer members at the time of their registration.

d. Ensure that registrants are counseled as to the conditions and requirements of their apprenticeship. If desired, request the expertise of an accessible Navy Campus education specialist for this counseling.

4. Commanding Officers of Registered Apprentices

a. Cancel the registration of personnel for any of the reasons listed in paragraph 204.1.

b. Suspend the registration of personnel for any of the reasons cited in paragraph 205.3.

c. Upon completion of all required hours of work experience, urge an apprentice to submit a request for issuance of a Certificate of Apprenticeship Completion.

d. Ensure that legitimate "hours completed" entries are made on the work experience hourly record in the work experience logs

of apprentices and that these entries are verified weekly by leading petty officers or work center supervisors.

e. Urge an apprentice to report twice a year, with work experience log, to the most accessible Navy Campus education specialist for an apprentice progress interview.

f. When desired, request the services of an accessible Navy Campus education specialist for the actions outlined in 4a, b, and c above.

5. Navy Campus Education Specialists

a. Provide maximum assistance, advice, and guidance to commanding officers, officers in charge, and directors of Class "A," "C," and "J" schools for the registration of eligible personnel and for the counseling of registrants as to the conditions and requirements of apprenticeships.

b. Provide maximum assistance, advice, and guidance to commanding officers of registered apprentices for the cancellation or suspension of registrations.

c. As permitted or requested by commanding officers, assist individual apprentices in requesting a Certificate of Apprenticeship Completion after all required hours of work experience have been completed. In preparing the final apprentice progress/status report, the Navy Campus education specialist will regard previous work experience as distributed proportionately among the skill areas of the applicable work processes schedule.

d. Be available for twice yearly apprentice progress interviews and reviews of work experience logs. The Navy Campus education specialist will inspect the work experience log of the apprentice and complete the appropriate blocks of the apprentice progress/status report. A professional advisement session will also be conducted, if appropriate.

e. Submit two copies of the apprentice progress/status report to the Chief of Naval Education and Training, together with one reproduced copy of any work experience hourly record which has been verified since the last apprentice progress interview. One copy of the apprentice progress/status report will be inserted in the work experience log of the apprentice.

f. Maintain modest stocks of forms required to administer subject program. Provide copies of required forms to apprentices and potential apprentices on an individual basis as requested.

6. Individual Apprentices

a. Request registration after successful completion of the applicable Class "A," "C," or "J" school.

b. At the time of registration, request the appropriate work experience log and thereafter assume responsibility for its safekeeping.

c. Request credit, if eligible, for previous work experience validated by page 4 service record entries or letters from previous employers. Assume responsibility for procurement of this documentation.

d. Enter completed hours of work experience in the work experience hourly record of the work experience log and have entries verified by the leading petty officer or work center supervisor.

e. Report twice yearly, with work experience log in hand, to the most accessible Navy Campus education specialist for an apprentice progress interview and submission of an apprentice progress/status report. Whenever possible, these interviews should be scheduled at least five months apart.

f. If operational requirements, hospitalization, assignment to duties not related to the trade in which registered, or inaccessibility of a Navy Campus education specialist prevent twice yearly progress interviews, request temporary suspension of registration in accordance with paragraph 205.

g. Upon completion of all hours of work experience required by the apprenticeship, report for a final apprentice progress/status interview. The final report will require verification of total hours of work experience in each skill area by commanding officer or accessible Navy Campus education specialist and submission of one reproduced copy of any work experience hourly record which has been verified since the last apprentice progress interview.

207. AVAILABILITY OF FORMS AND REQUIRED REPORTS

1. Sample copies of work experience logs for apprenticeable trades will be provided to Navy Campus education specialists by the Chief of Naval Education and Training Support. Only commanding officers, officers in charge, and directors of Class "A," "C," or "J" schools providing related instruction for designated apprenticeable trades are authorized to issue work experience logs to registered apprentices. Available work experience logs are listed in Appendix A.

2. Modest supplies of the following forms will be provided to Navy Campus education specialists and to Class "A," "C," and "J" schools providing related instruction for designated apprenticeable trades:

Apprentice Registration Application, CNET Form 1560/1
Apprentice Progress/Status Report, CNET Form 1560/2
Work Experience Hourly Record, CNET Form 1560/3.

3. Replacement supplies of the forms listed in the above paragraphs may be obtained by letter request to the Chief of Naval Education and Training Support, Pensacola, FL 32509. The forms will not be provided directly to individuals or to commands other than those cited above. Individuals having a need for any of the forms will obtain them from the most accessible Navy Campus education specialist.

4. Reports required for the National Apprenticeship Program have been approved by the Chief of Naval Operations. Report symbols apply as follows:

Report Symbol	Title	Form
CNET Report 1560-4	Apprentice Registration Application	CNET Form 1560/1
CNET Report 1560-5	Apprentice Progress/Status Report	CNET Form 1560/2.

208. AUTHORIZED TRADES. The Chief of Naval Education and Training and the Department of Labor have authorized the following apprenticeable trades within the active duty Navy:

1. Office Machines Mechanic

a. Term of apprenticeship: 6000 hours of specified and recorded work experience.

b. Dictionary of Occupational Titles (DOT) Code: 633.281.034.

c. All required related instruction satisfied by completion of any three of the listed packaged courses at the Service Schools Command, Naval Training Center, Great Lakes, IL 60088, provided that the applicant has previously completed the Instrumentman, Class "A" course. Personnel who are not graduates of the Instrumentman, Class "A" course are required to complete any five of the six courses in order to qualify for registration.

A-670-0028 IBM Selectric Typewriter Repair
A-670-0029 IBM C-1 Electric Typewriter Repair

A-670-0031	Remington Adder, Model 4, Repair
A-670-0032	Friden Calculator, Model STW, Repair
A-670-0034	Electronic Calculator Repair
A-670-0045	Bell and Howell Ditto Combomatic Copier/Duplicator, Model 441N, Repair.

d. Source Rating: Instrumentman (IM).

e. NEC's which can be translated into credit for previous work experience: None.

2. Watch-Clock Repairer

a. Term of apprenticeship: 6000 hours of specified and recorded work experience.

b. DOT Code: 715.281.034.

c. All required related instruction satisfied by completion of Watch Repair Course, Instrumentman Class "C1" (A-670-0011), Service Schools Command, Naval Training Center, Great Lakes, IL 60088.

d. Source Rating: Instrumentman (IM).

e. NEC's which can be translated into credit for previous work experience on a full year basis up to a maximum of 3000 hours: IM-1812.

3. Commercial Photographer

a. Term of apprenticeship: 6000 hours of specified and recorded work experience.

b. DOT Code: 143.062.034.

c. All required related instruction satisfied by completion of Photographer's Mate School, Class "A1" (Level 1) (C-400-2011) or Photographer's Mate School, Class "C7" (Level 2) (C-400-2012), Naval Technical Training Center, Corry Station Detachment (Photo School), Naval Air Station, Pensacola, FL 32508.

d. Source Rating: Photographer's Mate (PH).

e. NEC's which can be translated into credit for previous work experience on a full year basis up to a maximum of 2000 hours: PH-8148.

4. Camera Repairer

a. Term of apprenticeship: 4000 hours of specified and recorded work experience.

b. DOT Code 714.281.014.

c. All required related instruction satisfied by completion of Photographic Equipment Repair Course, Class "C1" (Level 3) (C-670-2012), Naval Technical Training Unit, Corry Station Detachment (Photo School), Naval Air Station, Pensacola, FL 32508.

d. Source Rating: Photographer's Mate (PH).

e. NEC's which can be translated into credit for previous work experience on a full year basis up to a maximum of 2000 hours: PH-8192.

5. Cook (Hotel and Restaurant)

a. Term of apprenticeship: 6000 hours of specified and recorded work experience.

b. DOT Code: 313.381.022.

c. All required related instruction is satisfied by completion of one of the following courses:

Mess Management Specialist Course, Food Production, Class "C1" (A-800-0018), Service Schools Command, Naval Training Center, San Diego, CA 92133. (Formerly Commissaryman/Steward, Class "C1" (A-800-0018), Food Production Course).

Mess Management Specialist Course, Management Principles, Class "C1" (A-800-0015), Service Schools Command, Naval Training Center, San Diego, CA 92133 or Fleet Training Center, Norfolk, VA 23511.

d. Source Rating: Mess Management Specialist (MS).

e. NEC's which can be translated into credit for previous work experience on a full year basis up to a maximum of 3000 hours: MS-3503; MS-3527; MS-3528; MS-3529; MS-3531; MS-3532; MS-3533.

APPRENTICE REGISTRATION APPLICATION
CNET Form 1560/1 (Rev. 11-75)

CNET Report 1560-6

1. Print or type.
2. Prepare in Triplicate.
3. Forward original and one copy to CNET.
4. Apprentice retains one copy in Work Experience Log.

PRIVACY ACT NOTIFICATION

Under the authority of 5 USC 301, the information regarding your former active military service, educational background and present personal data is requested in order to review and evaluate your qualifications for the Department of Labor apprenticeship program for active-duty Navy personnel. Your social security number will be used for purposes of individual identification. This information will be retained by CNET (Code N-11) and by the Bureau of Apprenticeship and Training, U. S. Department of Labor, and will not be divulged without your written authorization to anyone other than CNET and DOL personnel involved with the administration of this program. You are not required to provide this information; however, failure to do so may result in your not being registered for an apprenticeable trade and will militate against the Navy Campus for Achievement to be able to provide an education and training service to you.

APPLICANT INFORMATION

1. Name (Last, first, middle): [illegible], John Hortense
2. SSN: 399-[illegible]-[illegible]83
3. Date of birth: [illegible]
4. Sex: [X] MALE [ ] FEMALE
5. Race/ethnic group: [ ] CAUCASIAN/WHITE [X] NEGRO/BLACK [ ] AMERICAN INDIAN [ ] SPANISH AMERICAN [ ] ORIENTAL [ ] INFORMATION NOT AVAILABLE [ ] NOT ELSEWHERE CLASSIFIED
6. Name and location of high school from which graduated: Washington High School, Milwaukee, WI
7. Did you serve on active duty on or after 5 August 1964 and before 8 May 1975? [X] YES [ ] NO
8. Home of record (state): Wisconsin
9. Apprenticeable trade for registration: Watch-Clock Repairer
10. DOT code for apprenticeable trade: 715.281.03[illegible]
11. Apprentice program: 9 9 9 9 0
12. Signature of applicant. I agree to report to an NCFA Educational Specialist within 3 to 6 months after date of this application and twice a year thereafter. I understand that my registration ... and my registration does not guarantee work or duty assignments upon date of my apprenticeship. [signature]
13. Date: 29 July 1976

To be filled in by applicant's Class "A" or Class "C" School

TO:
Chief of Naval Education and Training
(Code N-11)
Naval Air Station
Pensacola, FL 32508

14. FROM
Director, CT/DK Schools
Service School Command
NTC, Great Lakes, Illinois 60088

15. Total hours required for term of apprenticeship: 6000 Hours
16. Hours credit given for previous work experience: (-) 1000 Hours
17. Total hours remaining for term of apprenticeship: 5000 Hours

18. Remarks: Registered by mail. [illegible] satisfactorily completed Watch Repair Course (A-670-0011) on 5 June 1975. [illegible] credited with 1000 hours of previous work experience for assignment to [illegible] for one year beginning 7 June 1975. Work Experience Log (CNET 1560/2) mailed on 17 Aug 1976 to: [illegible] John H. [illegible], USN, USS [illegible] (AD-26), R-5 Division, FPO New York, NY 09501.

19. Signature of official: [signature] [illegible], [illegible]
20. Title: Assistant Director, CT/DK Schools
21. Date: 17 Aug 1976

29 July 1976

From: IM2 John H. MELOWIG, USN, 399-64-6483
USS SHENANDOAH (AD-26) R-5 Division, FPO New York 09501
To: Director, OM/IM Schools, Service Schools Command, Great Lakes, Illinois 60088
Via: Commanding Officer

Subj: Registration in National Apprenticeship Program; request for

Encl: (1) Apprentice Registration Application (triplicate)
(2) Copy of service record, page 4

1. Enclosure (1) is forwarded for completion of registration. I completed the Watch Repair Course, Instrumentman Class "C1" (A-670-0011) on 5 June 1975 and have therefore satisfied the related instruction requirements for the Watch-Clock Repairer apprenticeship.

2. I request credit for 1000 hours of previous work experience based upon my assignment to the Navy Enlisted Classification IM-1812 since 7 June 1975. Enclosure (2) is forwarded in verification of this NEC.

3. It is further requested that the Work Experience Log for the Watch-Clock Repairer apprenticeship be forwarded to me so that I may begin to record my work experience on a regular basis.

JOHN H. MELOWIG

- -

30 July 1976

FIRST ENDORSEMENT on IM2 John H. MELOWIG ltr of 29 Jul 1976

From: Commanding Officer, USS SHENANDOAH (AD-26)
To: Commanding Officer, Service Schools Command (Director, OM/IM Schools), Great Lakes, Illinois 60088

Subj: Registration in National Apprenticeship Program; request for

1. Forwarded, recommending approval.

L.M. KOUSELIK
By direction

EXHIBIT II-B

Enclosure (1)

APPRENTICE PROGRESS/STATUS REPORT

CNET Form 1560/2 (4-75) — CNET Report 1560-5

1. *Print or Type*
2. *Prepare in Triplicate*
3. *Forward original and one copy to CNET with attached photo of last Hourly Record of Work Experience.*
4. *Apprentice retains one copy in Work Experience Log.*

PRIVACY ACT NOTIFICATION

Under the authority of 5 USC 301, the information regarding your former active military service, educational background and present personal data is requested for purposes of individual identification. This information will be retained by CNET (Code N-11) and by the Bureau of Apprenticeship and Training, U. S. Department of Labor, and will not be divulged without your written authorization to anyone other than CNET and DOL personnel involved with the administration of the apprenticeship program. You are not required to provide this information; however, failure to do so may result in cancellation of your registration in an apprenticeable trade.

To be filled in by Apprentice or official in accordance with instructions on reverse side.

1. Name of Apprentice (Last, First, Middle): M[illegible]TO, John Hortense	2. S.S.N.: 399-64-6483	3. Sex: ☒ Male ☐ Female

4. Race/Ethnic: ☐ Caucasian/White ☒ Negro/Black ☐ American Indian ☐ Spanish American ☐ Oriental ☐ Information Not Available ☐ Not Elsewhere Classified

5. Did You Serve On Active Duty On Or After 5 August 1964 And Before 8 May 1975? ☒ Yes ☐ No		6. Name of Record (Rating): Misconds
7. Apprenticeable Trade in Which Registered: Watch-Clock Repairer	8. Total Hours for Term: 6000	9. Hrs Prerequisite Experience: 1000
10. Hrs Completed Since Registration: 752	11. Hours Remaining: 4244	

ACTION REQUESTED

(Check One)

12. Please suspend registration for the apprentice named above for the reason(s) checked below:

a. ☐ [illegible]
b. ☐ [illegible]
c. ☐ Failure of current assignment probably work in apprenticeable trade for one year or less.
d. ☐ Operational commitments prevent reporting for progress interview or [illegible] Specialist not accessible.

13. ☒ Please lift the suspension of registration for the apprentice named above effective 9 February 1977
Date

14. ☐ Please cancel the registration of the apprentice named above for the reason(s) checked below:

a. ☐ Rated in [illegible] of professional competence
b. ☐ Discharge or release to inactive duty
c. ☐ Termination of work experience for one year or more
d. ☐ Death
e. ☐ Failure to report for twice-a-year apprentice progress interview
f. ☐ Personal request of apprentice

15. The apprentice named above has completed all required hours of work experience in all areas of the apprentice trade. A "Certificate of Apprenticeship Completion" is requested.

16. Signature of Apprentice: John Melonio — John H. [illegible]	17. Date: 9 Feb 1977
18. Signature and Title of Official: Fra[illegible], Navy Campus Education Specialist	19. Date: 9 Feb 1977

20. FROM

Navy Campus for Achievement
Building 2036
Naval Training Center
Orlando, FL 32813

TO:

Chief of Naval Education and Training
(Code N-11)
Naval Air Station
Pensacola, FL 32508

THE COMMUNITY COLLEGE OF THE AIR FORCE: TRADITIONAL TECHNICAL EDUCATION IN THE NON-TRADITIONAL MODE

Maj William A. Wojciechowski

OVERVIEW: Education program development within the Community College of the Air Force (CCAF) is consistent with its philosophy and in line with trends in both Air Force and civilian occupational education. Simply stated, CCAF intends to focus the educational effort of airmen toward obtaining a high quality, occupationally oriented career education which will enable them to serve as a master of and a supervisor in their specialty, either in the Air Force or in civilian life.

The U.S. Office of Education has estimated that 80% of the work force will require occupational skills in the 1980's as contrasted to baccalaureate skills. Concurrent with this changing emphasis on needed educational skills, we have seen the Air Force enlisted force decline from more than 800,000 personnel to slightly more than 500,000 in less than 5 years. With this reduction has come an expansion of technology requiring our personnel to be familiar with more complex weapon systems. Moreover, we see a growing need for more sophisticated supervisory skills. Past incidents of notable unrest indicate that knowledge of technology alone is inadequate--skill in human relations acquired through an understanding of the social sciences and the humanities is necessary. Therefore, the Community College of the Air Force has systematically plotted the career education of Air Force enlisted personnel to meet the demands of changing technologies and to develop the social awareness necessary for the effective management of people.

This paper depicts a model for the continuous development of CCAF programs and is a scenario describing the CCAF model in action. It explains how occupational requirements are translated into needs and interacted with resources to form the basis from which CCAF programs are developed. Discussion follows on the process of quality control as it is served by the CCAF Policy Council and the consultant advisory panel for each program. After review and approval of the program, articulation between CCAF, the Education Services Officer and the airman is defined. Finally, follow-up is discussed in terms of a feedback loop which provides continuity to the entire process.

OCCUPATIONAL REQUIREMENTS: Requirements for specific jobs

in the Air Force are determined from a review of programming documents, such as the Personnel Manpower Change Program series, which indicate numbers of personnel required in specific occupational specialties. The documents also reflect how many airmen, and with what training, leave the Air Force annually. The quality of airmen needed is indicated in Air Force Manual 39-1, Airman Classification Manual, and through means such as job inventory programs conducted by the Occupational Measurement Center of the Air Training Command, the Air Force Human Resources Laboratory, and feedback from using agencies.

Since it has been estimated by the Aerospace Education Foundation and the Air Force Human Resources Laboratory that from 60% to 90% of Air Force occupations have civilian counterparts, we believe a review of civilian occupational requirements is necessary as part of our program development process. The Department of Labor's Occupational Outlook Handbook and comparable state guides are reviewed to determine projected requirements for specific occupations. Long range projections are provided by the Department of Labor's Industry-Occupational Matrices and the President's manpower report.

An assessment of the quality of education required is then made through a review of guidelines such as those found in the USOE Technical Education Program series and Occupational Criteria and Preparatory Curriculum Plans in Technical Education Programs. Curriculum guides such as those for aviation and electro-mechanical careers published by the American Association of Community and Junior Colleges are also beneficial. The ASEE* Engineering Technology Education Study has been useful as a means of defining parameters of education for technicians. Guidelines prepared by other agencies such as the International Association of Firefighters and the Texas Commission on Law Enforcement Education have been studied and incorporated wherever possible to provide the necessary prerequisites for future licensing and certification of CCAF graduates.

RESOURCES: Prior to the existence of the CCAF's educational programs for enlisted persons, little emphasis was given to tying together service provided instruction and the related technical and general education available to the airman through a variety of sources. The Air Force provides technical training via some 2,500 to 4,000 technical training

* American Society for Engineering Education

courses to approximately 300,000 students per year. Approximately 80,000 students attend resident technical courses upon entering the Air Force. In addition, airmen receive management instruction through a series of NCO Academies and specialized instruction from agencies such as the School of Aerospace Medicine. Also, there is an extensive system for providing work experience combined with on-job instruction through a dual-channel OJT program--work and study. Completion of this period of apprenticeship is documented attesting to an individual's ability level to perform skills necessary in a particular specialty. This supervised training is identified as CCAF's Internship program for which credit is awarded.

Most of the aforementioned instruction is specifically designed to prepare airmen either as technicians or as managers/leaders. CCAF integrates this instruction with courses in related general education available from almost 400 civilian institutions which are associated with our 172 Base Education Services Centers. Included in CCAF's programs as an option is limited credit by examination offered through DANTES (Formerly the Armed Forces Institute (USAFI)).

CCAF CONTROL: If these myriad forms of instruction are to be focused toward career relevant education for airmen, program control must be exercised by a central agency. CCAF serves this function for the Air Force. The Program Development Division of CCAF is charged with the responsibility for analyzing service instruction to determine which parts of Air Force instruction:

- Are at a civilian post-secondary level. (This evaluation is made by program administrators, analysts, and department heads who have an average of 18 years experience in their occupational specialties as well as undergraduate and graduate degrees).

- Have civilian applicability and/or are occupationally related including that which is exclusively Air Force oriented. Subject matter which meets these basic criteria are evaluated on the basis of an average of 30 contact hours of instruction being equivalent to one semester hour. CCAF credit is applicable toward a Community College of the Air Force Associate in Applied Science (AAS) Degree.

Using service instruction as a core and guidelines established by USOE and other standard setting agencies for 2-year occupationally oriented associate level programs, a basic career pattern for the AAS was established. Within a total minimum length of 64 semester hours, sub-minima of 24 semester hours technical education related to an airman's Air Force occupation, 25 semester hours of general education designed for personal enrichment and to enhance supervisory skills, and 6 semester hours of management instruction make up the basic program pattern. Many technical electives and general education courses in the program are obtained from accredited institutions by the airman during his off-duty time, either with the Air Force providing 75 percent tuition assistance or with VA assistance. Currently, only 24 hours of credit by examination may be used toward this requirement.

As is clear from the foregoing, we are interested in focusing students' efforts on education related to their occupational specialties within the Air Force in order to maximize job performance and to enhance opportunities for post service employment. Therefore, after outlining 64 hour minimum program patterns, the Program Development Division has reviewed all Air Force occupational specialties and clustered them into 85 programs of study in 5 major career areas, e.g., in Management and Logistics there are programs in general business, business management, and computer science. In the electro-mechanical career areas, some 17 programs were developed. Each of these was constructed using the best available curriculum guidance. For example, U.S. Office of Education guidance for technical education indicates that programs should provide students with a facility in mathematics, physical science principles relative to technical skills, and ability in communications skills along with the knowledge of a particular occupational specialty. To this core we have added a facility in social sciences, humanities, and management to insure that airmen attaining supervisory positions develop a social awareness necessary to coping with the increasingly complex demands on personnel in management positions.

After the 85 programs were initially developed, they were reviewed by the CCAF Policy Council (annual review), chaired by the Dean and consisting of members of the College staff to include experts in the areas being reviewed. The programs were reviewed against criteria such as the following:

- o Does the program provide for Air Force occupational needs?

- o Does the program have a civilian occupational orientation?

- o Does the program meet the established minimum of 64 semester hours?

- o Does the program satisfy minimum criteria for accreditation such as those expressed in the Southern Association's Commission on Colleges guidelines?

After modifications suggested by the Policy Council to the programs are made, advisory panels of consultants are brought into action.

ADVISORY PANELS: To provide assurance that programs developed and controlled by CCAF are consistent with Air Force needs and, where possible, civilian requirements, we have identified advisory panels of consultants. They are representatives from the technical schools which offer instruction reflected in the programs; representatives from the Air Training Command Technical Training Directorate or Surgeon General; and representatives from business, education or industry. Wherever possible, representatives from appropriate professional organizations are asked to review CCAF programs and comment on them in terms of adequacy in preparing individuals to fulfill professional duties at the technician level. If licensing or certifying agencies exist for an occupational specialty, members of those agencies are also asked to review appropriate programs to determine their adequacy in preparing students for certification and licensing. Finally, in those areas where employment entry is dominated by unions or specific industries, members of the unions or industry may also be asked to comment. As an example, CCAF is interested in how its programs meet apprentice requirements. Although these advisory panels cannot be formalized because of federal regulations, informal contacts are continuously occurring. Consultants representing the groups noted above are employed and formally review CCAF programs. Their reports are available for examination.

PROGRAM DISTRIBUTION/GUIDANCE: Subsequent to annual review by advisory panels, programs are modified as necessary by the staff of the Program Development Division and formally approved by the Policy Council. Thereafter, the programs

will be incorporated into the CCAF catalog for distribution to Air Force schools and bases, recruiters, high schools, Education Services Officers, college registrars, and requesting employers. The catalog clearly indicates who can enter a CCAF program. It also provides specific guidance to Education Services Officers concerning CCAF programs and what courses airmen should be advised to take to progress through a degree program. The catalog clearly specifies how an airman may request a transcript of his technical education completed while in the Air Force and how he may have an official transcript forwarded to employers or colleges. Finally, the catalog details the procedure by which an airman may accumulate the necessary semester hours for an AAS and submit documentation to CCAF in support of his request for the degree.

ASSOCIATE IN APPLIED SCIENCE DEGREE APPROVAL: Once the airman's documentation is received at CCAF (after review and consolidation by Education Services Officers), it is reviewed by program administrators to insure that courses completed are from accredited institutions and are consistent with program objectives and that the airman has a coherent body of knowledge reflecting a comprehensive grasp of his occupational specialty. Airmen who fulfill AAS requirements will be recommended to the Policy Council for approval and award of the Associate in Applied Science degree.

FOLLOW-UP: To assure that the degree programs continue to meet the needs of the Air Force and, where possible, civilian employers, follow-up studies are conducted by the Institutional Research Branch of the CCAF Registrar. The studies determine how CCAF programs contribute to the improvement of NCO quality in terms of vocational skills and supervisory competence and how useful the programs are to employers and to other agencies such as colleges and universities. Feedback from these studies, as well as informal feedback from registrars, surveys, and other sources, enable us to modify and improve the programs as necessary.

SUMMARY: The systematic model which illustrates CCAF program development is educationally sound. Occupational requirements are illustrative of needs. Educational resources are available to fill these needs and CCAF programs result. Quality control in terms of substance and amount occur as a result of subsequent reviews from the CCAF Policy Council and advisory panels. Necessary modifications result and, then, final approval. The resulting career programs are published in the CCAF catalog. Interaction between CCAF,

the airman and his Education Services Officer occurs. This results in continuous vocational guidance with the objective being receipt of the AAS. Subsequent follow-up of CCAF graduates is made to determine how well CCAF programs have prepared the individual for his role as a technician and supervisor. The follow-up serves as a feedback loop making the CCAF program development process a continuous one.

DANTES Discussion Panel

ARMY SKILL DOCUMENTATION PROGRAMS

Lt Col Hal W. Downey

1. Soldiers receive, in varying degrees, training and experience in skills which have counterparts in the civilian community. As a result, many soldiers leaving the Army are proficient in skills needed by civilian industry and could be assimilated into the economy without extensive training. Experience has shown that industry has not, as a general rule, accepted exservicemembers as skilled workers. Because this attitude is costly to industry we must assume it is based upon a lack of awareness of the value of military training and experience rather than a reluctance to employ veterans at an advanced starting level. The Army, in an effort to create an awareness of the value of military training and experience in civilian industry, in the early 1970s began to develop programs to obtain recognition from selected industries for the skills of soldiers in certain Engineer, Transportation and Culinary specialities. Pilot programs in apprenticeship, industry recognition of non-apprenticeable skills, and technical certification were initiated.

2. The initial programs were developed by Training and Doctrine Command (TRADOC) Service Schools, as the technical expertise in the skills the programs represented was in the training committees of these schools. There was no central Army control and there were no uniform procedures for program development. These latter facts caused concern at Headquarters, Department of the Army and in late 1974 The Adjutant General was given the responsibility for coordinating program development and for prescribing uniform policies for program development and management. At this time the three pilot programs were well underway.

a. The National Apprenticeship standards for the Army were developed, and in July 1975 were registered officially with the Bureau of Apprenticeship and Training (BAT), US Department of Labor. The Engineer School Apprenticeship Program, developed jointly by the Department of Labor, the US Army Engineer School and three civilian agencies (the International Union of Operating Engineers, the Associated General Contractors of America, and the National Constructors Association) was registered with BAT in August 1975 and became the Army's first apprenticeship program. This program, consisting of four occupational skills (Plant Equipment Operator, Grading and Paving Equipment Operator, Universal Equipment Operator and Heavy Duty Repairer (Construction Equipment)) was implemented Army-wide in February 1976. From the initial pilot program in 1973 to implementation of an Army program in 1976, we learned a great deal about apprenticeship programs, much of it by trial and error. The prolonged developmental period for these first programs shortened the overall development of the Army apprenticeship program.

b. The Transportation School developed their pilot program to gain recognition for enlisted personnel who desire a career in the fields of motor transport operations, marine craft operations and maintenance, aviation maintenance, or terminal operations and whose military experience and training has been in there fields. This program was developed in coordination with some 40 national and regional industries which employ people in the aforementioned career fields. Industry representatives, after reviewing training programs of the Transportation School, indorsed the Transportation Corps Industry Accreditation Program (IAP). The IAP is essentially a referral program thru which soldiers about to leave the Army prepare a brief, formated resume of their Army experience, list areas of desired employment, and are referred to employers who could provide them jobs. This program will serve as the model for the manner of developing Industry Recognition Programs in coordination with interested national/regional industries; however, we do not anticipate development of any further referral programs.

c. The Quartermaster School developed a pilot program in the area of professional/technical certification for food service personnel. They contacted national/regional agencies which were leaders in the culinary field to invite them to participate in development of an Army program. Only the American Culinary Federation (ACF) had a certification program and desired to work with the Army. The resultant program was one which outlined procedures for soldiers to follow to gain ACF certification at the various levels of culinary expertise and provided for participants to join ACF chapters. After much debate it was finally determined that the Army would not sponsor any private association programs and since ACF was a private association the pilot was not to become an Army program. Instead, the Quartermaster School was to develop an Industry Recognition Program for food service personnel as a follow-up to their cooks' apprenticeship program or, an alternative form of documentation for cooks.

3. The Army set as its goals, the development of programs to document skills attained by soldiers, in a form which would have meaning to civilian industry. There would be no alteration of Army training or of duty assignments simply to provide skills useful to soldiers in seeking post-service employment; there would be no programs developed to train soldiers in civilian skills; there would be no programs designed primarily to gain certification from private associations.

a. The Army determined to target its initial programs on enlisted personnel, as they comprise the bulk of the soldiers leaving the service.

b. Because apprenticeship programs offer a well structured and meaningful form of documentation of progression in skills, we elected to concentrate first on Army skills which had counterpart civilian apprenticeable skills and then to move on to other directly relatable skills or to programs beyond apprenticeship, for apprenticeable skills.

c. We determined to develop some experimental programs to attempt to relate some of the general skills of personnel in specialities which have no direct civilian counterparts (e.g. combat arms) to those experiences employers would consider attributes in prospective employees (i.e., leadership, administration, stocking, etc.)

d. We will also, either using the catalog developed by the Defense Activity for Non-Traditional Education Support (DANTES) or thru our own research, identify skills requiring licensing/certification as a prerequisite for employment and indicate the regulating agencies.

4. Army Apprenticeship Programs for Military Personnel.

a. These are programs paralleling those in the civilian community, developed for Army skills which relate to civilian apprenticeable skills, registered with the US Department of Labor (DOL), thru which participants can achieve DOL certification as Journeyman.

b. The TRADOC Service Schools who have training proponency (responsible for the development of training programs to qualify personnel in the skills of the MOS) for an enlisted speciality (MOS) design apprenticeship programs with the assistance of the Bureau of Apprenticeship and Training (BAT), DOL. They send their programs thru command channels to BAT requesting the programs be registered under the National Standards of the Army and the apprenticeship standards of their school. After a program is registered the school prepares a draft of a DA Pamphlet which will implement the program Army-wide. Once a program is implemented, the school acts as technical advisor to Education Services Officers and insures programs are kept current.

c. Army Apprenticeship Programs are a part of the Army Continuing Education System. They are a means of professional development for a soldier and are operated thru Army Education Centers. Soldiers register with their Education Services Officer who in turn sends registration data to the Army Adjutant General Center (TAGCEN) where data is stored and given each month to BAT.

d. Soldiers have individual logbooks which contain cummulative records of work experience, validated by their supervisors. Each apprenticeship program has a schedule of work processes in which an apprentice must log a specific number of hours of satisfactory work in order to reach Journeyman status. Programs vary depending upon the complexity of the skill from a total 2000 hours of experience to 8000 hours and for each 2000 hours there is a corresponding requirement that the apprentice obtain 144 hours of related instruction (in many cases Advanced Individual Training will satisfy the full program requirement for instruction).

e. If a soldier logs all required hours of work experience and attends the requisite hours of related instruction he/she may apply to the local Education Services Officer (ESO) for a Certificate of Completion of

Apprenticeship. These certificates will be issued by DOL. The ESO simply verifies the logbook entries of the individual and, once the entitlement is established, requests that TAGCEN obtain a certificate from DOL. TAGCEN sends the signed certificates to the ESO for appropriate presentation.

f. Soldiers who leave the Army after having completed only a portion of the total required hours of experience and/or related instruction present their logbooks to the ESO for verification. The ESO will then issue a letter certifying the number of hours completed in each work process and attesting to the fact the individual was participating in a Nationally registered apprenticeship program.

g. The Army had 29 programs implemented as of 30 September 1977, 30 more registered awaiting publication of DA Pamphlets for implementation, and 13 at BAT under review. This constitutes the initial group of programs to be developed. Additional programs will be developed as new areas are opened to apprenticeship (e.g. a law enforcement program was recently registered with BAT and copies of this program, obtained by TAGCEN in Sep 77, were furnished to the Military Policy School for consideration). See Incl 1 for program listing.

h. We believe that apprenticeship programs will offer soldiers an appealing means of professional development with a clearly defined personal goal. As soldiers pursue this goal they will become more professional in their MOS and, because most all programs will take more than one enlistment to complete, there will be an additional incentive for reenlistment for a 2d term of service. As the value of these programs can be documented (from surveys of soldiers leaving the service and from evaluations by participants) they should present a valuable recruiting tool. The Army benefits will accrue as soldiers strive toward their individual goals.

5. Industry Recognition Programs (IRP)

a. These will be programs developed in conjunction with employing industries, to establish a form of credential which will adequately define for potential employers an individual's level of skill proficiency, either in a non-apprenticeable skill or that achieved beyond journeyman status in an apprenticeable skill.

b. Current plans call for the development of IRP's to be completed by September 1978. The Veterans Employment Service of DOL has agreed to assist the Army in developing IRP's.

6. In summary:

a. We know soldiers possess valuable skills, attained thru military training and experience, which in many cases relate to occupational skills in civilian industry.

b. Our programs are designed to document this skill proficiency in a form which has practical significance to potential employers.

c. We do not alter training or assignment to qualify an individual in a civilian skill, nor do we run special training programs which detract from the time a soldier spends in readiness training.

d. We do not promise post-service employment nor are we in job placement. We offer a voluntary means of professional development, as part of the Army Continuing Education System.

IMPLEMENTED ARMY APPRENTICESHIP PROGRAMS

ENGINEER SCHOOL

PROGRAM TITLE	LENGTH
Plant Equipment Operator	6000 hours
Grading & Paving Equipment Operator	"
Heavy Duty Repairer (Const Equip)	"
Universal Equipment Operator (Const Equip)	"

QUARTERMASTER SCHOOL

PROGRAM TITLE	LENGTH
Cook	6000 hours
Laboratory Technician (petroleum)	"

TRANSPORTATION SCHOOL

PROGRAM TITLE	LENGTH
Sheet Metal Worker (Aircraft)	6000 hours
Electrical Mechanic (Acrft)	"
Marine Heavy Duty Mechanic (Hvy Dty Mech - Diesel)	8000 hours
Marine Hull Repairer, Ironworker (Boatbuilder, Steel)	"
Airplane Mechanic	6000 hours

SIGNAL SCHOOL

PROGRAM TITLE	LENGTH
Radio Communications Technician	8000 hours
Aircraft Electrical Mechanic	"
Central Office Telephone Installer & Repairer	"
Automatic Equipment Technician	"
Radio/TV Repairer	"
Cable Splicer	"

ORDNANCE SCHOOL

PROGRAM TITLE	LENGTH
Small Weapons Repairer	8000 hours
Artillery Repairer	"
Industrial Welder	6000 hours
Machinist	8000 hours
Automobile Body Repairer & Painter	"
Sewing Machine Repairer	6000 hours
Automobile Mechanic	8000 hours
Truck Mechanic	"
Heavy Duty Equipment Mechanic	"

IMPLEMENTED ARMY APPRENTICESHIP PROGRAMS (Continued)

MISSILE/MUNITIONS SCHOOL

PROGRAM TITLE	LENGTH
Electronics Technician (Radar)	7000 hours
Electrical Instrument Repairer	6000 hours
Hydraulic-Equipment Mechanic	7000 hours

(Inclosure 1)

ATTITUDINAL CORRELATES OF REENLISTMENT INTENT AMONG WOMEN IN THE ARMY

Jack M. Hicks

U. S. Army Research Institute for the
Behavioral and Social Sciences

INTRODUCTION

The investigation upon which this report was based, grew out of a need for the U. S. Army to better understand the contributions which young women might potentially make to its manpower system. This need derived in particular from the fact that women constitute 50% of the general population, and thus greatly expand the Army's potential manpower source. The overall objective of this effort was to generate a better understanding of the host of variables pertaining to the adjustment of enlisted women (EW) in the Army, including current attitudes, morale, perceptions, motivations, and reenlistment intent.

The present report was generally responsive to current Army concerns with respect to retention of first-term enlisted personnel. More specifically, this analysis addressed the currently little studied domain of reenlistment intent correlates among Army EW. The purpose of this report was to identify some of the most promising correlates of self-reported reenlistment intent among those attitudinal variables included in the above discussed parent investigation. The long-range objective of this research is to develop a profile of reliable correlates of reenlistment intent, specifically applicable to EW. It is hoped that ultimately, such an analysis might serve as a basis for an EW reenlistment predictor battery.

METHOD

Based upon preliminary interviews and a pilot study, a 162-item questionnaire was designed. Nine division-size Army installations with a high percentage of EW were sampled for data collection, yielding a total of 1718 EW respondents. To the extent possible, selection of EW was governed by the following specifications: (1) length of service, 18-24 months (first-term); (2) at least 9 months from expiration of term of service (ETS); (3) primarily E-3 and E-4; (4) eligible for reenlistment; and (5) diversity of MOS.

The questionnaire was administered by civilian data collection teams to collectives of 25 to 100 EW per session, in December 1975. Confidentiality and anonymity were assured. The verbatim instructions were read aloud, with the respondents following silently from their questionnaire booklets. A one hour period proved sufficient to complete the questionnaire.

RESULTS

Twenty-five percent of the EW indicated that they would definitely or probably reenlist at ETS. An additional 26% indicated that they might or might not reenlist, with the remaining 49% reporting that they would definitely or probably not reenlist.

Enlistment Considerations. The interpretation of relationships between initial enlistment motivations and reenlistment intent should take into account that a minimum of 18 months had passed since enlistment, and that ETS was still perhaps 18 months in the future. Such time lags might be expected to spuriously deflate whatever true relationships exist. Nonetheless, several statistically significant correlations resulted. The statistical tests employed were chi square to establish significance/nonsignificance, and Cramer's V to estimate strength of relationship. Of the 11 most frequently reported enlistment motivations, 5 were found to be significantly associated with reenlistment intent. Of some interest was the fact that the most frequently reported enlistment motives were not those which proved to be most predictive of reenlistment intent. In fact, the most popular enlistment motive, "to get college benefits," was not significantly correlated with reenlistment intent. The next most often given reasons for enlistment, "to get civilian job training," "to travel," and "to find adventure and excitement," were significantly, but modestly associated with reenlistment intent. The strongest correlates of reenlistment intent came from less often reported motivations such as, "to serve my country" (V = .21), and "came from a military family" (V = .16).

Also worthy of mention is the significant correlation between "length of time spent thinking about enlistment" and reenlistment intent (V = .24). In this case, those who indicated a liklihood of reenlistment, were significantly less inclined to say that they "joined on impulse," than those who reported little propensity to reenlist.

Satisfaction with the Army. The strongest association found in the entire analysis was between "overall satisfaction with the Army" and reenlistment intent (V = .56). Approximately half of those inclined to reenlist expressed high satisfaction with the Army overall, as compared to only 6% among those disinclined to reenlist. Only 3% of those with reenlistment plans indicated low satisfaction with the Army.

Almost as strong a relationship was yielded between "work satisfaction in the Army" and reenlistment intent (V = .47). As many as 81% of those intending to reenlist reported good to excellent work satisfaction. Only 39% of those leaning against reenlistment reported a similar level of work satisfaction. This finding was in direct corroboration of the primary finding in a very recent report on job satisfaction and reenlistment intent among enlisted men (Goldman and Worstine, 1977).

Interest in Combat. A major current issue regarding the utilization of women in the military concerns their potential for combat MOS,

from which they are excluded at present. One of the strongest correlations obtained in this analysis was that between "interest in combat" and reenlistment intent (V = .35). Approximately one half of those women interested in reenlistment also expressed an interest in combat. By contrast, less one in five of those not intending to reenlist reported a similar interest in combat. This finding is symptomatic of a general pattern inherent in this body of data, which suggests that women prone to reenlist are more inclined to feel that they can and should do men's work, than women not interested in reenlistment.

Perceived Quality of Personnel. Among the strongest correlates of reenlistment intent for enlisted women, was their perception of the quality and desirability of their male counterparts. Approximately 1/3 of the women intending to reenlist reported the overall quality of the men to be good to excellent, as compared to 13% for those planning not to reup. (V = .27). Almost one-half of those not intending to reenlist considered the men fair to poor, as compared to less that 1/4 of those planning to reenlist who felt this way. A similar result was obtained in connection with the perceived liklihood of marrying an Army man (assuming single status). Fully 1/2 of the reenlistment prone respondents indicated that they would be inclined to marry an Army man, as compared to approximately 1/4 of those not reenlistment prone (V = .27). Much the same results were obtained from the inclination to date an Army man, as related to reenlistment intent (V = .22), and the perceived overall quality of women in the Army, as related to reenlistment intent (V = .21).

Likes and Dislikes. Finally, the respondents were asked to indicate the things they liked and disliked most about the Army. In terms of relationships to reenlistment intent, the strongest correlate among these stimuli was "Army tradition" (V = .32). Almost 2/3 of the reenlistment prone respondents reported a liking for Army tradition, as compared to less than 1/3 for those not reenlistment prone. A similar relationship (V = .32) was obtained between "rules and regulations" and reenlistment intent. Other significant correlates of reenlistment intent were "field duty" (V = .30), "dress/hair regulations" (V = .26), "the NCO's" (V = .21), "dress uniforms" (V = .20), "the food" (V = .20), "comissioned officers" (V = .19), "this Post" (V = .10), "my MOS" (V = .18), and "fatigue uniforms" (V = .18). In all of these instances, the greater the liking, the greater the expressed intent to reenlist.

Somewhat consistent with the above findings concerning quality of personnel, liking of "most men in the Army" correlated more highly with reenlistment intent (V = .16) than liking of "most women in the Army" (V = .03). In both cases, however, the V's were much lower than found in regard to quality of personnel. It was also of interest that the relationship of "heavy lifting" and reenlistment intent, though statistically significant (V = .15), was not among the strongest predictors of reenlistment intent. Other poor correlates were like/dislike of "the barracks" (V = .10) and "abuse from civilians" (V = .04).

In summary, there appear to be, on the basis of this analysis a substantial number of promising variables which might serve as a basis for a first-term female reenlistment predictor battery. A key assumption relevant to this objective, of course, is a high relationship between reenlistment intent and actual reenlistment behavior. On the basis of previous research old and new, such an assumption would appear to be reasonably sound. According to the most recent research, the Air Force Human Resources Laboratory (Guinn, Berberich, and Vitola, 1977) found that 92% of those first-termers expressing an interest in a military career actually did reenlist; whereas, 93% of those who expressed disinterest, did not reenlist. Goldman and Worstine (1977) also reported that "...soldiers' statements regarding reenlistment intent are highly correlated with actual reenlistment decision..." Of course, these relationships tend to hold most strongly the less the time lag between expressed reenlistment intent and actual reenlistment. Nonetheless, there appears to be sufficient justification for the assumption that a high relationship exists between reenlistment intent and actual reenlistment such that the present analysis might serve as a useful preliminary exploration of the feasibility a predictor of reenlistment decision for women in the Army.

REFERENCES

Guinn, N. Berberich, G. & Vitola, B.M. Reenlistee/non-reenlistee profiles and prediction of reenlistment potential (AFHRL-TR-77-24). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, Personnel Research Division, June 1977.

Goldman, A.C. & Worstine, D.A. Job satisfaction and reenlistment intent for first term personnel: initial findings (Survey Report). Alexandria, VA: US Army Military Personnel Center, Personnel Management Systems Directorate, May 1977.

APPENDIX

TABLE 1

TERM OF SERVICE AND REENLISTMENT INTENT

<u>RE-UP INTENT</u>

TERM	WILL	MAYBE	WONT	TOTALS
FIRST	373	407	773	1553 (91%)
NOT FIRST	44	36	73	153 (9%)
TOTALS	417 (25%)	443 (26%)	846 (49%)	1706

CHI SQUARE = 1.78

$P > .05$

TABLE 2

MOTIVATIONS AND REENLISTMENT INTENT

RE-UP INTENT

ENLISTMENT MOTIVATION	WILL	WONT	P	V
GET COLLEGE BENEFITS	54%	57%	NS	--
GET CIVILIAN JOB TRAINING	53%	42%	< .0001	.103
TRAVEL	50%	41%	< .01	.077
ADVENTURE EXCITEMENT	48%	37%	< .001	.105
NEEDED THE MONEY	25%	25%	NS	--
NO CIVILIAN JOB	17%	20%	NS	--
SERVE COUNTRY	27%	11%	< .001	.206
GET AWAY FROM HOME	13%	17%	NS	--
GET AWAY FROM SMALL TOWN	13%	13%	NS	--
DIDN'T LIKE JOB	11%	12%	NS	--
CAME FROM MILITARY FAMILY	13%	5%	< .001	.156

OVERALL CHI SQUARE = 77

$p < .001$

TABLE 3

LENGTH OF TIME THOUGHT ABOUT ENLISTING

AS RELATED TO REENLISTMENT INTENT

	RE-UP INTENT	
LENGTH OF TIME	WILL	WONT
NOT LONG, JOINED ON IMPULSE	21%	36%
THOUGHT ABOUT IT SEVERAL MONTHS	32%	36%
THOUGHT ABOUT IT A YEAR OR SO	24%	21%
WANTED IN ARMY SINCE VERY YOUNG	22%	7%

CHI SQUARE = 70

$P < .001$

$V = .235$

TABLE 4

OVERALL SATISFACTION WITH ARMY

AS RELATED TO REENLISTMENT INTENT

	RE-UP INTENT	
SATISFACTION	WILL	WONT
HIGH	49%	6%
MEDIUM	48%	50%
LOW	3%	43%

CHI SQUARE = 396

$P < .001$

$V = .559$

TABLE 5

WORK SATISFACTION IN ARMY

AS RELATED TO REENLISTMENT INTENT

	RE-UP INTENT	
WORK SATISFACTION	WILL	WONT
EXCELLENT	38%	8%
GOOD	43%	31%
AVERAGE	16%	31%
FAIR	2%	15%
POOR	1%	14%

CHI SQUARE = 26

$P < .001$

V = .468

TABLE 6

INTEREST IN COMBAT AS RELATED

TO REENLISTMENT INTENT

	RE-UP INTENT	
INTEREST IN COMBAT	WILL	WONT
A LOT	23%	6%
SOME	26%	12%
NOT VERY MUCH	21%	17%
NOT AT ALL	30%	64%

CHI SQUARE = 158

$p < .001$

V = .354

TABLE 7

PERCEIVED CHANCES OF LEARNING CIVILIAN JOB SKILL

AS RELATED TO REENLISTMENT INTENT

	RE-UP INTENT	
CHANCES OF SKILL	WILL	WONT
EXCELLENT	33%	14%
GOOD	41%	29%
AVERAGE	18%	24%
FAIR	4%	12%
POOR	4%	21%

CHI SQUARE = 145

$p < .001$

V = .339

TABLE 8

PERCEIVED OVERALL QUALITY OF MEN AS RELATED TO

REENLISTMENT INTENT

	RE-UP INTENT	
QUALITY	WILL	WONT
EXCELLENT	3%	2%
GOOD	29%	11%
AVERAGE	43%	39%
FAIR	16%	27%
POOR	8%	21%

CHI SQUARE = 94

$P < .001$

$V = .274$

TABLE 9

PERCEIVED LIKLIHOOD OF MARRYING AN ARMY MAN IF SINGLE;

AS RELATED TO REENLISTMENT INTENT

	RE-UP INTENT	
MARRY ARMY MAN	WILL	WONT
DEFINITELY WOULD	23%	13%
PROBABLY WOULD	28%	13%
MIGHT/MIGHT NOT	30%	36%
PROBABLY WOULDN'T	8%	16%
DEFINITELY WOULDN'T	11%	21%

CHI SQUARE = 89

$P < .001$

$V = .266$

TABLE 10

PERCEIVED LIKLIHOOD OF DATING AN ARMY MAN IF

SINGLE, AS RELATED TO REENLISTMENT INTENT

	RE-UP INTENT	
DATE ARMY MAN	WILL	WONT
DEFINITELY WOULD	34%	21%
PROPABLY WOULD	37%	28%
MIGHT/MIGHT NOT	21%	34%
PROBABLY WOULDN'T	4%	8%
DEFINTELY WOULDN'T	3%	9%

CHI SQUARE = 63

$P < .001$

V = .223

TABLE 11

PERCEIVED OVERALL QUALITY OF WOMEN AS RELATED TO

REENLISTMENT INTENT

	RE-UP INTENT	
QUALITY	WILL	WONT
EXCELLENT	6%	2%
GOOD	36%	22%
AVERAGE	38%	44%
FAIR	15%	22%
POOR	4%	10%

CHI SQUARE = 58

$P < .001$

V = .214

TABLE 12

LIKES AND DISLIKES ABOUT THE ARMY AS RELATED TO REENLISTMENT INTENT

STIMULUS		RE-UP INTENT WILL	RE-UP INTENT WONT	
ARMY "TRADITION"	DISLIKE	26%	55%	V = .325
	LIKE	63%	32%	
RULES AND REGS	DISLIKE	29%	58%	V = .322
	LIKE	58%	28%	
FIELD DUTY	DISLIKE	44%	71%	V = .296
	LIKE	45%	20%	
DRESS/HAIR REGS	DISLIKE	44%	68%	V = .257
	LIKE	45%	22%	
THE NCO'S	DISLIKE	13%	29%	V = .206
	LIKE	74%	55%	
DRESS UNIFORMS	DISLIKE	28%	45%	V = .205
	LIKE	63%	41%	
THE FOOD	DISLIKE	47%	65%	V = .205
	LIKE	44%	25%	
OFFICERS	DISLIKE	17%	35%	V = .194
	LIKE	69%	51%	
THIS POST	DISLIKE	46%	62%	V = .186
	LIKE	47%	28%	
MY MOS	DISLIKE	18%	33%	V = .185
	LIKE	75%	55%	
FATIGUES	DISLIKE	34%	51%	V = .183
	LIKE	59%	40%	

MALE/FEMALE ATTITUDES RELATED TO PERFORMANCE IN AIR FORCE TECHNICAL TRAINING

Jeffrey E. Kantor, Bart M. Vitola, and Nancy Guinn
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

At the request of the USAF Air Training Command, the Personnel Research Division of the Air Force Human Resources Laboratory initiated a research project designed to investigate student attitudes toward Air Force technical training and the relationship between those attitudes and performance/attrition in technical training. This research was divided into three main phases: (1) the development and validation of an instrument both sensitive to student attitudes and related to technical training performance, (2) a comparison of student attitudes from courses having differential attrition rates, and (3) an assessment of the relationship between attitude and performance in general and attitude and performance within specific subgroups of interest.

The first phase, development and validation of the Technical Training Student Survey (TTSS) was completed in early 1977 (Kantor, Vitola, & Guinn, 1977), and the remaining phases are, at present, nearing completion. In the course of this effort, a data base was established consisting of attitudinal responses and technical training course performance measures on 12,667 technical training students. From this data base, it is possible to abstract and study various subgroups of interest. Attitudinal differences between these groups can be identified and relationships between

attitude and performance can be compared. In this study, the subgroups chosen for comparison were male and female students. Comparisons drawn between males and females are of interest for several reasons. First, male/female differences have, historically, been an area of both scientific and popular inquiry. Second, with the increase in numbers of women entering the Air Force, and the military in general, it has become important to identify and assess male/female differences which might impact on personnel training and utilization. Finally, in many technical training areas, particularly Mechanical and Electronic, males and females exhibit differential attrition rates unrelated to entering aptitude scores. Therefore, the objectives of this study were to (a) identify attitudinal differences between male and female students regarding Air Force technical training and (b) compare and contrast the relationships between attitudes and performance for male and female technical training students.

Method

Subjects. A total of 12,667 nonprior service enlisted accessions (10,980 males and 1,687 females) were administered the TTSS while attending one of 53 different Air Force technical training courses conducted between September 1974 and August 1975. For comparative purposes, this sample was subdivided, based upon sex and performance in technical school, into four groups: (1) Male Graduates (9,984), (2) Male Eliminees (996), (3) Female Graduates (1,430), and (4) Female Eliminees (257).

Survey Instrument. The TTSS contains 121 items designed to tap student attitudes about specific aspects of the Air Force technical

training experience. These measures reflect the student's expectations about training; motivation for training; perceptions of instructors, fellow students, and physical settings; degree of perceived stress in training; and the degree of personal satisfaction derived from the student's training and career choice. Approximate testing time for the TTSS is 30 minutes. A copy of the TTSS is presented in Appendix A. An example of the type of item and response format used is presented in Figure 1.

Survey Administration. The TTSS was administered under standardized conditions to students in the training setting. Sampling points were chosen to allow comparisons across all technical training courses, between technical training centers, and between courses having differing attrition rates. It is assumed that the response patterns obtained accurately reflected the spectrum of attitudes present in the population of Air Force technical training students.

Statistical Analysis. To evaluate male/female and graduate/eliminee differences, a stepwise discriminant analysis approach was utilized. This technique provided both an identification of specific differential attitudes and a relative importance weighting of those differences. Since only two groups were compared at any one time, this approach is analogous to a multiple linear regression with dichotomously coded dependent variables. Error rate (Type I) was controlled per family of stepwise comparisons such that the total α for each set of comparisons $\leq .05$.

Results and Discussion

The first analysis was accomplished to identify attitudinal differences between male and female students. For this analysis, sex was the dependent variable and significant relationships were identified between the sex of the respondent and his or her responses on 33 of the 121 items. These 33 items accounted for 9.48% of the dependent variance. Based upon the item content and the relative weight of that item in the discriminant function, the major attitudinal differences between males and females were summarized and are presented in descending order of importance in Figure 2. (A complete list of the 33 items and their correlations with the dependent variable is presented in Appendix B.1.)

From these attitudinal differences, a few general findings seem apparent. Females show more concern about academics (i.e., desire more off-duty study time, desire more time be spent on difficult subject matter). This is possibly related to their higher attrition rate from technical training schools (males = 8.98%; females = 15.23%) but may reflect a desire to perform up to standards even if additional time and effort are required. Females are less satisfied with certain aspects of the physical environment (classroom temperature, dorm sleeping facilities) but have a more positive perception of their fellow students (less petty quarrels, more support). Finally, although females seem more happy with their military status (more satisfied with the Air Force, less bothered by military bearing), it is the males who felt that technical training had been a more beneficial experience. Overall, it appears that the females

evidenced more academic difficulty, more group cohesion, more satisfaction, but perhaps were less sure of what benefit they were getting out of training. These attitudes might be considered typical of those of a group entering into a new environment, and it is possible that as the numbers and experiences of females in technical training increase, some of the male/female differences will be moderated and the similarities increased.

To differentiate between the attitudes of male graduates and male eliminees, an analysis was accomplished using the 10,980 male subjects with graduation/elimination being the dependent variable. Significant relationships were identified between the dependent variable and responses on 22 of the 121 TTSS items accounting for 9.76% of the dependent variance. (A complete list of these items is provided in Appendix B.2.) The major attitudinal differences between male graduates and eliminees are summarized in Figure 3.

From these attitudinal differences, it would appear that male eliminees felt more stress (pressure for perfection, difficulty with materials, interference with studies), that male graduates placed more importance on system rewards (job security, avoidance of duties), and that both male graduates and eliminees held some negative feelings about each other. Overall, it might be that the male eliminee evidences more susceptability to pressure, less personal motivation, and less affinity for inherent system reinforcers. This makes the eliminee easily discouraged and very difficult to keep on track and working when arduous effort is required.

To differentiate between the attitudes of female graduates and eliminees, an analysis was accomplished using the 1,687 female subjects again with graduation/elimination being the dependent variable. Significant relationships were identified on 12 of the 121 items, accounting for 11.52% of the dependent variance. The major attitudinal differences between female graduates and eliminees are summarized in Figure 4. (A complete list of the 12 items is presented in Appendix B.3.)

From these attitudinal differences, it would appear that female eliminees also felt more stress (pressure for perfection, difficulty with course materials, student workload), that female graduates were more motivated (desire more study time, more time on equipment), and that female graduates placed more importance on system rewards (job security, off-duty privileges). Again, like the males, it would appear that the female eliminee evidences more susceptability to pressure, less drive towards the goal, and might be difficult to motivate since she appears less sensitive to system reinforcers.

The major attitudinal factors found related to graduation/elimination for males and females are summarized and compared in Table 1. It would appear evident that considerable overlap exists between the factors associated with technical training performance for males and females. Out of the first five most important factors, four are shared by both males and females, leading to the conclusion that the similarities outweigh the differences between the sexes. However, the differences which do exist appear to point to the conclusion that females have somewhat

more academic difficulty than males. Since all students entering any particular training course are qualified for that course and have generally comparable aptitude scores, this finding is interesting because it suggests a difference in ability not currently being measured. Several areas of future research are therefore suggested. First, it should be determined if the relationships between aptitude test scores and performance in technical school are the same for both males and females. Second, course materials and structure should be investigated for sex bias which might negatively impact on female performance. Finally, the Air Force selection and classification system, developed on a primarily all male force, should be evaluated to ensure that females are being properly managed with respect to the maximally effective classification of female personnel and their assignment to areas wherein they will have the highest probability of success.

Conclusion

The male and female attitudes regarding the Air Force technical training experience were found to differ significantly in several areas. Some of these differences (e.g., classroom temperature) may be dealt with directly, but most appear to be reflecting the differences in attitudes between a group with experience in a particular environment (males) versus those of a group entering a new experience (females). It is possible that as the ''newness'' of having large numbers of females in technical training wears off, the similarities between male and female students will increase. The similarities between factors associated with graduation/elimination for males and females are substantial and appear to indicate

similar problems in eliminees of both sexes. However, some differences were noted and appear to be indicative of females having more academic difficulties. In summary, certain attitudinal differences do exist between males and females in Air Force technical training, but there was substantial communality indicating similar perceptions, concerns, and a similar relationship between attitude and performance.

Reference

Kartor, J.E., Vitola, B.M., and Guinn, N. Development and Validation of the Air Force Technical Training Student Survey. AFHRL-TR-77-27(1). Brooks AFB, Texas. Personnel Research Division, Air Force Human Resources Laboratory, June 1977.

Figure 1. An Example of the Type of Item and Response Format Used in the Technical Training Student Survey

	Definitely Disagree				Definitely Agree
Certain students are hostile toward other class members					
Most students get along well together					
Fellow students look out for each other					
Certain students are uncooperative					
Certain students are responsible for petty quarrels and bad feelings among class members					
There are tensions among some students which interfere with training activities					
Certain students are incapable of working together					
Students help each other to learn the necessary course material					
Some students are not liked or accepted by fellow students					
Students have to take advantage of others in order to succeed in training					

Note. Instructions for Responding:

1. Below is a series of statements related to both your training and training environment.
2. Please darken the one circle on each scale that best expresses your feelings.

II. Scale 4: Instructor Task Competence[c]

27. Ineffective — Effective
28. Knowledgeable — Ignorant
29. Boring — Interesting
30. Dependable — Undependable
31. Disorganized — Organized
32. Unsure — Confident
33. Convincing — Unconvincing
34. Unprepared — Prepared
35. Intelligent — Stupid
36. Inefficient — Efficient

[c]Instructions for Scales 4 and 5

1. Please use the scales below to describe your main (lead) instructor.
2. Darken the *one circle on each scale* that best expresses your feelings.

III. Scale 5: Instructor Personal Relations

37. Encourages — Discourages
38. Criticizes — Praises
39. Fair — Unfair
40. Impatient — Patient
41. Considerate — Inconsiderate
42. Hinders — Helps
43. Friendly — Unfriendly
44. Supportive — Hostile
45. Ridicules — Compliments
46. Cooperative — Uncooperative

Figure 2

MAJOR ATTITUDE DIFFERENCES BETWEEN MALES AND FEMALES

0 FEMALES DESIRE MORE OFF-DUTY STUDY TIME

0 FEMALES DO NOT THINK CLASSROOM TEMPERATURE IS SATISFACTORY

0 MALES SEE MORE PETTY QUARRELS AMONG FELLOW STUDENTS

0 MALES BELIEVE MILITARY BEARING DISTRACTS FROM SCHOOL PERFORMANCE

0 FEMALES DESIRE BETTER DORM SLEEPING FACILITIES

0 FEMALES BELIEVE STUDENTS LOOK OUT FOR EACH OTHER

0 FEMALES DO NOT THINK ENOUGH TIME IS SPENT ON DIFFICULT SUBJECTS

0 FEMALES ARE MORE SATISFIED WITH THE AIR FORCE

0 MALES THINK TECH TRAINING HAS BEEN MORE BENEFICIAL TO THEIR CAREER

Figure 3

MAJOR ATTITUDINAL DIFFERENCES BETWEEN MALE GRADUATES/ELIMINEES

- ELIMINEES FEEL MORE PRESSURE FOR PERFECTION
- JOB SECURITY MORE IMPORTANT TO GRADUATES
- AVOIDING "MICKEY MOUSE" DUTIES MORE IMPORTANT TO GRADUATES
- ELIMINEES BELIEVE THEY CAN COMPLETE TRAINING AHEAD OF SCHEDULE
- ELIMINEES BELIEVE COURSE MATERIALS ARE TOO HARD
- GRADUATES THINK CERTAIN STUDENTS ARE HOSTILE
- ELIMINEES SEE CERTAIN TENSION BETWEEN STUDENTS
- ELIMINEES BELIEVE SQUADRON DUTIES INTERFERE WITH STUDIES
- ELIMINEES THINK INSTRUCTORS ARE BORING
- GRADUATES WANT MORE TIME ON TRAINING EQUIPMENT

Figure 4

MAJOR ATTITUDINAL DIFFERENCES BETWEEN FEMALE GRADUATES/ELIMINEES

0 ELIMINEES FEEL MORE PRESSURE FOR PERFECTION

0 GRADUATES DESIRE MORE OFF-DUTY STUDY TIME

0 ELIMINEES BELIEVE COURSE MATERIALS ARE TOO HARD

0 JOB SECURITY MORE IMPORTANT TO GRADUATES

0 ELIMINEES BELIEVE THEY CAN COMPLETE TRAINING AHEAD OF SCHEDULE

0 ELIMINEES BELIEVE STUDENT WORKLOAD IS TOO HEAVY

0 GRADUATES DESIRE MORE TIME ON TRAINING EQUIPMENT

0 OFF DUTY PRIVILEGES MORE IMPORTANT TO GRADUATES

Appendix A

TECHNICAL TRAINING STUDENT SURVEY
PE 7403
AIR FORCE HUMAN RESOURCES LABORATORY

SSAN

GENERAL INSTRUCTIONS:

1. The items contained on this form are designed to measure student attitudes toward Air Force technical training.
2. The form is intended to give you the opportunity to help improve student training.
3. It is very important that your answers reflect your true feelings. This is not a test and you are not required to put your name on the form.
4. Please carefully follow the instructions at the beginning of each of the four main sections of this form.

DATE: Yr. 19__ Da. Mo. JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

SECTION I:

1. Below are statements describing rewards a student might receive if he performs well in technical training.
2. Beside each statement are two separate rating scales.

   On Scale 1 indicate how likely it is for you to receive the reward if you perform well in training.

   On Scale 2 indicate how important the reward is to you. Consider only its importance, not how likely or unlikely you are to receive the reward.
3. Notice that each scale has five circles. The words above the scales describe the meaning of the circles at the ends of each scale. The three circles in the middle of each scale represent feelings between those described at the scale ends. You might want to think of each scale as similar to a thermometer lying on its side.
4. Answer each item by darkening one circle on each scale to indicate how you feel about the statement. Read each statement carefully and take all the time you need.

	SCALE 1 IF YOU PERFORM WELL (Very Unlikely … Very Likely)	SCALE 2 HOW IMPORTANT TO YOU (Not Important … Very Important)
1 Increased job security after graduating from technical school	○ ○ ○ ○ ○	○ ○ ○ ○ ○
2 Faster promotion	○ ○ ○ ○ ○	○ ○ ○ ○ ○
3 Greater chance to participate in important decisions after graduating from technical school	○ ○ ○ ○ ○	○ ○ ○ ○ ○
4 More challenging duty assignments after graduating from technical school	○ ○ ○ ○ ○	○ ○ ○ ○ ○
5 More job responsibilities after graduating from technical school	○ ○ ○ ○ ○	○ ○ ○ ○ ○
6 Greater chance of being skilled and competent in your career field	○ ○ ○ ○ ○	○ ○ ○ ○ ○
7 Increased chance of getting a good civilian job after Air Force service	○ ○ ○ ○ ○	○ ○ ○ ○ ○
8 Greater chance to be assigned to your base of choice	○ ○ ○ ○ ○	○ ○ ○ ○ ○
9 Increased off-duty privileges (for example three-day passes or no squadron detail)	○ ○ ○ ○ ○	○ ○ ○ ○ ○
10 Greater freedom in deciding how to accomplish class work	○ ○ ○ ○ ○	○ ○ ○ ○ ○
11 Increased chance of being admired and respected by fellow students	○ ○ ○ ○ ○	○ ○ ○ ○ ○

ATC Form 1631 Jun 74 — GO TO THE NEXT PAGE

PRINTED BY NRC IOWA CITY IOWA FORM NO 3883 W 2330

Note. Option Coding: 1 2 3 4 5 (Scale 1); 1 2 3 4 5 (Scale 2)

SECTION I:

	SCALE 1: IF YOU PERFORM WELL		SCALE 2: HOW IMPORTANT TO YOU	
	Very Unlikely	Very Likely	Not Important	Very Important
12. Instructors pay more attention to your ideas and suggestions				
13. Increased educational growth and development				
14. Greater chance to help other students learn the subject matter				
15. Greater chance to do better on tests and receive better grades				
16. Receive compliments, recognition and praise from instructors				
17. Fewer "Mickey Mouse" duties in the Squadron				
18. Fewer "Mickey Mouse" assignments in class				
19. Feeling of self respect and sense of accomplishment				
20. Increased opportunity to use your abilities				
21. Receive more challenging class assignments				
22. Greater opportunity to study subject matter of special interest to you				
23. Increased chance of completing training ahead of schedule				
24. Provided with more spare time				
25. Instructors less critical of your work				
26. Increased chance of being an "Honor" graduate				

SECTION II:

1. Please use the scales below to describe your SSAN of main (lead) instructor
2. Darken the one circle on each scale that best expresses your feelings.

27	Ineffective	Effective
28	Knowledgeable	Ignorant
29	Boring	Interesting
30	Dependable	Undependable
31	Disorganized	Organized
32	Unsure	Confident
33	Convincing	Unconvincing
34	Unprepared	Prepared
35	Intelligent	Stupid
36	Inefficient	Efficient
37	Encourages	Discourages
38	Criticizes	Praises
39	Fair	Unfair
40	Impatient	Patient
41	Considerate	Inconsiderate
42	Hinders	Helps
43	Friendly	Unfriendly
44	Supportive	Hostile
45	Ridicules	Compliments
46	Cooperative	Uncooperative

GO TO THE NEXT PAGE

SECTION III

1. Below are a series of statements related to both your training and training environment

2. Please darken the one circle on each scale that best expresses your feelings.

SCAN

	Definitely Disagree				Definitely Agree
47 Certain students are hostile toward other class members					
48. Most students get along well together					
49. Fellow students look out for each other					
50. Certain students are uncooperative					
51 Certain students are responsible for petty quarrels and bad feelings among class members					
52 There are tensions among some students which interfere with training activities					
53. Certain students are incapable of working together					
54. Students help each other to learn the necessary course material					
55. Some students are not liked or accepted by fellow students					
56. Students have to take advantage of others in order to succeed in training					
57 Students are given an equal opportunity to demonstrate their capabilities					
58. Students are subject to strict discipline					
59 Student training is too closely supervised					
60 Students are encouraged to speak their minds even if it means disagreeing with the instructors					
61 Students are encouraged to suggest improvements or solutions to training problems					
62 Students are encouraged to participate in classroom discussions					
63. Students are given the opportunity to participate in class					
64. Student suggestions and recommendations are considered with fairness					
65. Students are seldom able to use their own judgment					
66. Students have no say about what happens to them					
67 Students have little chance to influence the way the class is conducted					
68. Students have the freedom to establish their own study schedules					
69 Spare time in class may be spent as each student sees fit					
70. Students are rarely given the chance to freely express their ideas in the classroom					

GO TO THE NEXT PAGE

SECTION III.

Definitely Disagree — Definitely Agree

71 Students are seldom allowed to act independently

72 Pressure for perfection is unbearable

73 The military atmosphere in the classroom interferes with learning of the subject matter

74 Squadron duties interfere with study

75 In order to do well in training, students have to do things that are against their personal values

76 Students don't know what is expected of them

77 There is confusion in the planning and organization of classroom activities

78 There is considerable conflict among training objectives

79 Performance standards are unreasonably high

80 Emphasis is placed on passing the course rather than learning subject matter

81 There is a good deal of disagreement on how this training should be conducted

82 The student work load is too heavy

83 The quantity of class work interferes with how well it is done

84 Emphasis on military bearing and appearance detract from student performance

85 Training hours are too long

86 Conflicts exist in the training requirements

87 Training equipment (including trainers) is adequate

88 Training equipment (including trainers) is readily available for student practice

89 Time allowed on training equipment (including trainers) is sufficient

90 Training evaluation or testing is an accurate indication of student performance

91 Study guides are difficult to understand

92 Excessive attention is given to unimportant details

93 Course materials are so poor that they contribute little to learning

94 Course materials are not closely related to the course objectives

95 Course materials are more difficult than they should be

96 My progress in class is not what it should be due to the poor quality of training or course materials

97 Classroom temperature is satisfactory

98 Dormitory sleeping facilities are adequate

GO TO THE NEXT PAGE

SSAN

SECTION III:

	Definitely Disagree				Definitely Agree
99 Classroom lighting is adequate					
100 Classroom chairs are comfortable					
101 Classroom seating arrangement is satisfactory					
102 Length of class breaks is about right					
103 Number of class breaks is sufficient					
104 Study facilities outside the classroom are adequate					
105 Classroom ventilation is about right					
106 Time allowed for testing is sufficient					
107 Classroom noise control is effective					
108 Classroom work space is sufficient (desk or table top area)					
109 Supplementary study materials (manuals, regulations, technical orders, etc.) are readily available for student use					
110 Base recreation facilities are adequate					
111 Off duty study time is sufficient					
112 Time allowed for review of tests is adequate					
113 Enough training time is spent on difficult and important subject matter					
114 Some students would perform better on a different shift					

SECTION IV

1 Below are statements about your satisfaction with your training and career field.

2 Please darken the circle that best expresses your feelings about the statement in the same way you have in the other sections of this form.

	Completely Dissatisfied				Completely Satisfied
115 How do you feel about your technical training					
116 How do you feel about your assigned career field?					
117 How do you feel about the Air Force?					

GO TO THE NEXT PAGE

SECTION IV:

	Completely Different				Identical
118 How similar is your assigned career field to your preferred career field?	⬭	⬭	⬭	⬭	⬭
	Highly Inaccurate				Highly Accurate
119 How accurate was the information you received about your career field before entering technical training?	⬭	⬭	⬭	⬭	⬭
	Strongly Negative				Strongly Positive
120 What effect has technical training had on your feelings about your career field?	⬭	⬭	⬭	⬭	⬭
	Definitely No				Definitely Yes
121 If you have the chance, will you change to another career field?	⬭	⬭	⬭	⬭	⬭

REMARKS:

APPENDIX B.1

TTSS items significantly related to sex of respondent coding: Males = 1, Females = 2: Item options coded as per Appendix A.

Item #	Correlation	Item #	Correlation
111	-.117	75	-.034
97	-.111	82	.064
51	-.075	104	-.079
84	-.049	66	-.025
98	-.104	4(2)	-.035
49	.059	119	-.005
113	-.095	112	-.090
117	.038	25(1)	.018
120	-.063	54	.050
109	.007	59	-.036
2(2)	-.054	62	-.030
19(1)	.029		
56	-.058		
115	-.078		
38	.025		
29	-.054		
88	.073		
110	-.069		
8(1)	.033		
69	-.054		
118	-.054		

Note. Items are listed in order of entry into the stepwise discriminate analysis.

APPENDIX B.2

TTSS items significantly related to graduation/elimination of male students coding: Graduation = 0, Elimination = 1.

Item #	Correlation
72	.182
1(2)	-.124
17(2)	-.081
23(1)	.074
95	.128
47	-.058
52	.061
74	-.020
29	-.091
88	.046
80	.099
110	.031
12(2)	-.003
79	.139
13(2)	-.082
51	-.032
70	.097
3(1)	.019
82	.109
89	.041
69	-.035
84	.031

Note. Items are listed in order of entry into the stepwise discriminate analysis.

APPENDIX B.3

TTSS items significantly related to graduation/elimination of female students coding: Graduates = 0, Eliminees = 1.

Item #	Correlation
72	.209
111	.086
95	.154
1(2)	-.117
23(1)	.121
82	.142
89	.049
9(2)	-.056
80	.134
84	.004
62	.009
33	.117

Note. Items are listed in order of entry into the stepwise discriminate analysis.

ARMY COMMANDERS AS GATEKEEPERS FOR INFORMATION

David L. Hannaman
Dr Robert Pulliam

ABSTRACT

The Army's Chief of Public Affairs performed a study of company-level command, to determine the effects of "gatekeepers" on the flow of command information from DA to soldiers in the field. Commanders and their staffs in 102 representative companies were surveyed, using a structured interview analyzed by Q-sort techniques. Findings confirmed the critical role of commanders and of other media in determining how soldiers perceive the Army and its mission, and revealed interesting characteristics of modern small-unit command. Commanders were found to be primary gatekeepers for only 30% of the information flow studied. Films appeared to be less effective as media than is commonly believed, and limits of enlisted reading skill appeared to be a critical constraint on choice of media.

The survey technique used is of technical interest because of its utility for further research in the Services.

INTRODUCTION

There has been some kind of a formal discipline in military psychology for about 50 years. Yet after 50 years we are not very close to an understanding of the dynamics of unit behavior nor of command. Alexander the Great probably knew more, intuitively, about the psychology of command than we yet have learned from research.

My colleagues and I therefore continuously search for research tools by which we can measure and describe the relationship between command and unit behavior. And to us, one of the most promising tools is the study of the flow of information. This applies equally to the formal flow of orders, regulations, or combat intelligence, and to the informal flow of private information and unofficial working communications. It applies equally to military units in combat, and to housekeeping duties in garrison.

We were therefore particularly pleased when the Army selected Kinton, Inc., to study the flow of command information at the company level. Today we will report that study briefly. We will describe the requirement, outline our method and report some highlights of the findings. We believe the methods used have wide applicability in military research for addressing other problems.

But, first, we wish to express our appreciation to the many persons in the Army who contributed to this research, and without whose assistance the study would not have been possible. First, we would like to recognize the support of Colonel Ralph E. Ropp and Captain Carroll W. Williams who recognized the requirement for this research and who contributed substantially to its design and direction. We are further indebted to the Public Affairs Officers at Forts Belvoir, Benning, Bragg, Dix, Hood, Lee,Meade and Polk, and to members of their staffs for untiring work in scheduling and arranging interviews. Several valuable observations were first offered by senior commanders and were later confirmed by company level interviews.

Finally, we are most deeply indebted to the hard-pressed company commanders and first sergeants who took time to tell us how the information program operates in the working Army. This study is properly dedicated to those professionals, and we hope it tells some small part of their story.

REQUIREMENT

The study we are reporting was planned by the Army as a means of evaluating the Army's Command Information program. When we say "Command Information" we refer to general information not directly related to the Army's technical mission, which senior commanders consider important to dissiminate down the line of command. It includes a wide range of matters, ranging from changes in personnel policy to the Army's position on weapons systems procurement. For convenience we will refer to Command Information as "CI".

The Army's CI Program provides specific information, identified as important by the Army, to individual officers and soldiers at all levels. When it works it should ensure that major policies and programs are understood, and that troops understand the Army's role and mission, as that role is seen by senior command. There are cases in which the view of the Department of Army is at variance with the attitudes of some soldiers (such as concerning race relations) or with popularly held attitudes (such as concerning the Soviet threat). In these cases the mission of the information program is to ensure that members of the Army at least understand the position the Army takes in pursuing its constitutional mission. Soldiers are not required to agree, but they need to understand the rationale for the Army's role.

Because of its importance, the Army is concerned about the effectiveness of the CI program, and has in the past undertaken studies of its effectiveness and of the comparative value of various media or vehicles for transmitting the program's messages. Research studies as early as World War II analyzed the "Why We Fight Fight" series, on separate scales for information, attitude and motivational effectiveness. A general finding of most studies has been that the information program is never fully effective in delivering information (cognitive content), but that the program is more effective in communicating information than in causing attitudinal or behavioral change.

Company Commanders

This specific study focused on the role of the company commander, as a "gatekeeper." We will say more about "gatekeepers" in a moment. There were reasons to believe that company commanders were the key to effectiveness in the Army CI Program. To begin with, they are in

most cases the point at which the CI program's materials distribution system stops, and decisions are made as to which materials will be presented, when, and how. The commander is the person responsible for formal Commander's Calls, and the one who must select what he will say to his troops during the limited time they are assembled together. Finally, he is the senior authority figure in the Army chain of command, the one who is most personally known and regularly seen by troops.

The company commander is therefore probably the gatekeeper whose decisions are most influential for the information program. There are two reservations in that regard:

First, the decisions made by the commander are often difficult to distinguish from those of the first sergeant and other orderly room staff - such decisions may either be based on the recommendations of others or may in fact normally bypass the commander. Many commanders do not actually see or read completely the information sent them through information program channels, but depend on others to screen those materials first.

Second, the tenure of commanders is sometimes brief, and the impact of the commander, when the turnover has been rapid, may be less than in those cases where the commander has been assigned long enough to establish his position and to develop an administrative routine.

Gatekeepers

The question we asked was about the commander's behavior as a "gatekeeper" of information. The role of gatekeepers was a key to the study. Gatekeepers determine what other people receive through channels of communication.

It is hard to fully appreciate how dependent each of us is on the content of communications. What we know, what we believe, and how we behave depends on the information we receive through formal communication channels. Each human being can know directly only those events which happen within his sight and hearing. All the rest of his perception of the world results from indirect experience, received through communications.

Thus most members of a modern society, including soldiers in the Army, have no direct experience of realities such as floods in Bangladesh, the space program, the Soviet Union, or the fact that the world is round. These things are observed by others, and communicated via

a host of private or public channels. Nevertheless, most people believe in the detailed existence of a wider world, most of which they have never seen.

How the world beyond personal experience is perceived is totally a function of communications. No citizen or soldier can know of any event, unless that event is reflected in the media he receives. How each person perceives an event (such as the war in Ethiopia) is totally dependent on the content of communications. That content is necessarily selective. The channels of communication could not carry, and individuals could not digest, all that happens in the world each day. So at many points in the world's communications, there are "gates" or "filters", points at which signals are sorted, edited and selected before being passed into the next channel. "Gatekeepers" are critical causes of the way the world is perceived by others.

Gatekeepers necessarily exist within the communications of society, the government, business, private institutions, and the Army. Either consciously or by default, gatekeepers determine what signals will be received by others, and therefore, how others will perceive reality. In the Army, commanders are the primary gatekeepers for information flowing with the chain of command. While they are not the only gatekeepers, the decisions of company commanders are centrally important in determining how troops and junior officers perceive the Army, and themselves, in relation to the Army's mission. This gatekeeping role is vital in battle as well as in peacetime affairs. It has never before to our knowledge, been the subject of specific research.

Finally, it must be observed that in the company commander's case, the term "gatekeeper" does not adequately suggest his information role. Gatekeepers in government, media and business often function in a manner closely analogous to the gate and filter functions of a computer - they pass or process information. But the company commander is in addition an active medium of transmission and display. Commanders are active advocates. How they promote a policy, and how they display their personal concern, will determine how their subordinates behave to a far greater extent than is the case elsewhere in government, business or industry.

We have explained that the objective of this study was to define the company commander's role as a gatekeeper of CI. We will now describe the methods of the study.

METHOD

In designing a method for this study we recognized a couple of problems to be overcome. One was that we did not yet know exactly what questions would be most useful to ask, and the other was that company commanders were likely to be on the defensive when asked about their CI programs. We had a good method for exploring the wording of questions in the "Q-sort" technique, but the question of defensiveness was more serious.

We know that military commanders are not likely to be frank when asked how they do things. They are used to more or less constant inspection, and habitually tell outsiders the story which will make the unit look good. But we needed to know what actually happened to CI at the company level, as contrasted to what the regulations or policy might say. We knew that most commanders would not be using all the CI materials fully, and that in fact commanders have to avoid many directed responsibilities in order to survive.

Structured Interview

Informality and personal contact were obviously required. We decided to use a structured interview, with open-ended responses, so that we could ask stimulating questions and then let the respondents tell us what was on their minds. The three interviewers were each former members of the military, with personal experience as a commander. This was our most valuable asset. But, first we had to decide what to ask.

The Questionnaire

The Army provided us with a list of questions which needed answering. These concerned the effectiveness of several publications, such as Soldiers magazine and Commander's Call, an information pamphlet. In addition, we were to explore the effectiveness of media such as radio, movies and TV, and to explore opinions at the company level, about the CI program. This list of general questions had to be pruned down to a practical number of concrete questions, and stated in terms which would provoke useful, measurable responses.

Case Study

A preliminary list of questions was drawn up, reflecting our best estimate of what questions would be useful and how they should be phrased. Those questions were tested in a case-study of 10 Army units. In each unit we interviewed the commander and his first sergeant; four units were in the combat arms, four were combat support and two were in Army schools.

The interviewers used approximately two hours in talking to each respondent. The questions were asked informally, in a conversational manner, and respondents were encouraged to answer informally and at length. Longhand notes were taken, in many cases recording the exact language of key phrases or sentences. New ideas and perceptions were picked up and explored further.

A principal purpose of these case studies was to discover how the issues were perceived at company level, so that our interview would reflect the language, issues and perceptions of unit-level command rather than those of Army headquarters.

Responses of commanders and their first sergeants were compared to detect unreliable data, and differences between enlisted and officer perception.

Q-Sort Scoring

A Q-sort procedure was used to identify scorable responses. By this we mean that a panel of analysts looked at each question, the analysts taking one question at a time and working independently. For each question, each analyst sorted the responses into categories of similar responses. A typical question will usually yield from two to eight recurring responses (plus a miscellaneous category). Sets of responses emerge which, though phrased differently, reflect similar opinions. The panel then meets, compares its scoring systems, and agrees on a common scoring scheme. This method makes it feasible to score well-designed, open-ended questions as if they were multiple-choice.

Model Description

From these data a preliminary logical gated-flow diagram was developed, to be confirmed and expanded later. That diagram described the possible patterns of flow for CI information within an Army company.

Mini Field Test

As a result of the case study, major modifications were made to the wording of questions and the question sequence. A preliminary field test was then conducted at Ft. Belvoir, Virginia. Following this test the procedures just outlined were repeated. Questions were scored, their language was improved, and minor adjustments made in format and sequence. A survey instrument resulted which was clearly in the language of users, was easy to Q-sort, which produced interesting data, and which was reliable to apply.

Field Test

This survey instrument then underwent a final rigorous field test. Twenty-nine units were surveyed at Forts Bragg, Dix and Lee. A total of 26 company commanders and 29 first sergeants were contacted, in 10 combat arms, 9 combat support, and 10 school companies. The data were scored and analyzed after which minor changes were made, but it was then determined that the field test data were sufficiently reliable to include in the final report.

Final Survey

A total of 102 units (61 combat arms, 21 combat support and 20 schools) from Forts Benning, Bragg, Hood and Polk were included in the final survey. Eighty-eight company commanders and eighty-four first sergeants were interviewed. Data from the Field Test and the final survey were consistent. However, the structured interview was changed in minor respects following the Field Test, so the data from the Field Test and Survey are not in every case fully comparable.

Each installation involved in the final survey was visited by two or more interviewers. This was required in part to lessen the probability of interviewer bias affecting the data. Additionally, the interviewers conducted at least two interviews jointly at each installation -- one interviewing while the other observed, then reversing roles for the second interview. This procedure ensured standardization of interviewing technique. Interviews were conducted in a company setting, typically the office, mess hall, training room or dayroom of the unit concerned.

Including the final survey, a total of 8 installations (Forts Belvoir, Benning, Bragg, Dix, Hood, Lee, Meade and Polk) and 144 units were visited during the course of the entire study. The total interview population included 127 COs and 124 first sergeants. Table 1 reflects the composition of this population by phases.

TABLE 1

Persons Interviewed, By Phase

Phase	Units	COs	1st SGs
Case Study	10	10	8
Preliminary Field Test	3	3	3
Field Test	29	26	29
Formal Survey	102	88	84
TOTALS	144	127	124

Data Analysis

All responses were evaluated using the Q-sort procedure. The data were posted to a master tabular matrix, and analyzed by inspection. Where the findings were interesting and significant, they were pulled out as summary tables. Table 2, for instance, reflects the opinions of commanders and of first sergeants as to whether enlisted men read Soldiers. It shows general agreement that Soldiers is widely read, but that readership is lower among less senior enlistees.

TABLE 2

Question: What percentage of your men, by grade, read Soldiers?

	Interviewees	
Grade	COs	1st SGS
E1-E4	48%	47%
E5-E6	69%	67%
PSGs	83%	88%
Platoon Leaders	83%	88%

A final diagram of information flow, as reported by respondents, was developed (Table 3). That flow was used further in reporting the findings.

FINDINGS

Our findings were reported to the Army in a series of briefings, and in a final report. Most of these have to do with individual Army publications and are not of general interest. We do want to report a few things which are of more general interest, and which probably apply in some measure to the Navy and Air Force, as well. We will limit ourselves to five points.

Gatekeeping and Command

In general, three dating behavior models were discerned. Table 3 applies. Those models are described following:

Commander is Prime Gatekeeper

In about 1/3 of all units, the CO did personally receive, review, and make principle decisions concerning actions to be taken on command information. In those cases, he would typically read incoming material selectively, evaluate its importance, and mark it for the attention of others and for further dissemination through formations, distribution or posting.

First Sergeant is Gatekeeper

In a slightly smaller percentage of units this function was performed by the first sergeant. Four situations were typical: (1) In more than 50% of cases the CO specifically delegated responsibility for reading and screening command information, and acted only on items called to his attention. (2) In other cases, the CO continued the practice which existed in the unit prior to his assignment, or otherwise found himself within a pattern of established unit behavior. This often involved Army "regulations" or policies and existing SOPs, directing the gating of information within the unit. This was perhaps the most interesting of the gating behavior models observed; it was identified early in the survey when one of the COs stated that there was a unit SOP for the handling and dissemination of CI. When he was asked who directed this SOP, he stated that he didn't know. "It was here and in effect when I got to the unit a year ago." When his first

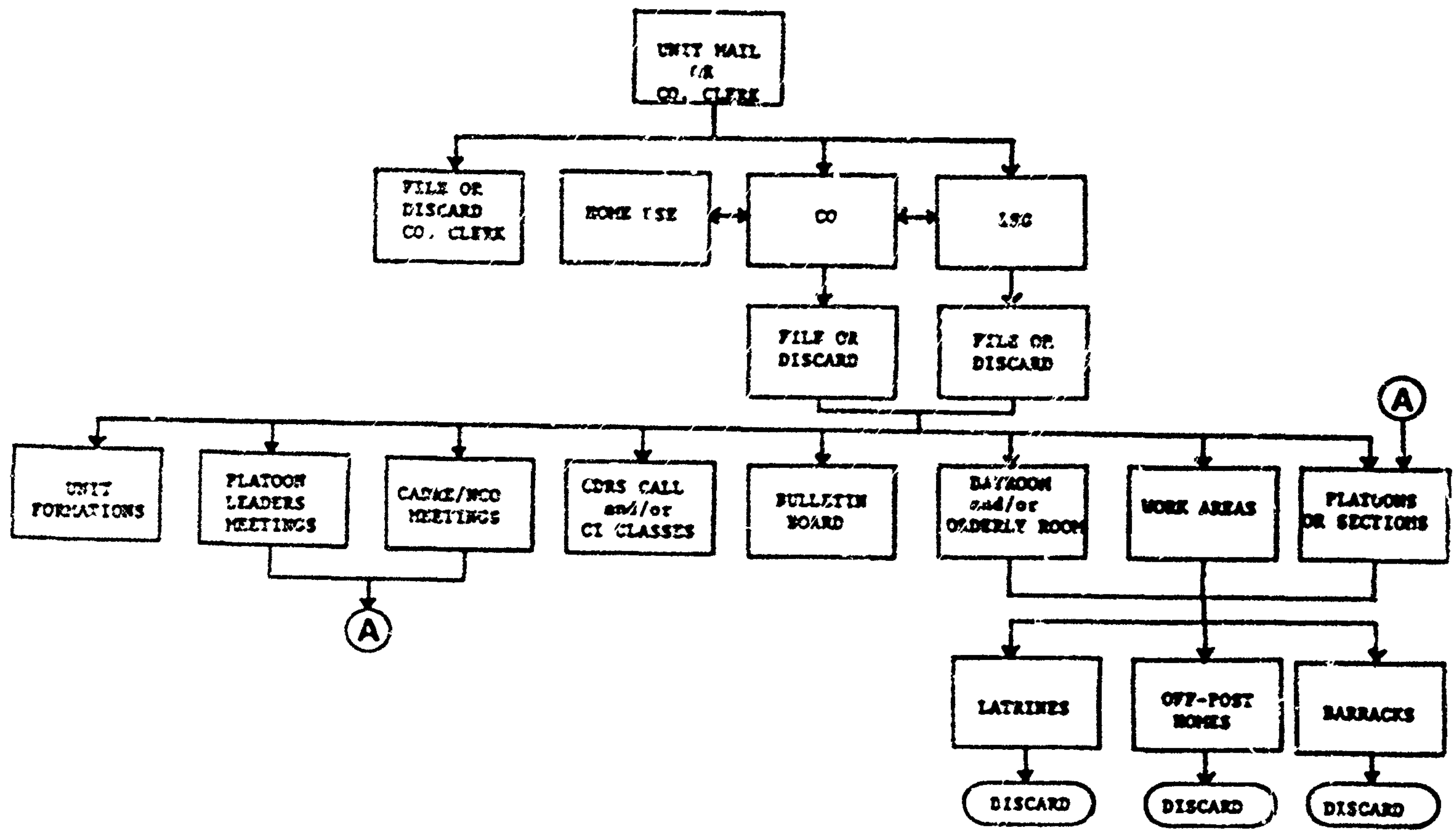


TABLE 3
COMPANY INFORMATION FLOW

sergeant was asked who directed the SOP, his response was identical to that of the CO with one difference: he had been the first sergeant of the unit for more than 13 months, and the SOP had been in effect when he arrived, too. Where CI dissemination SOPs did exist in units, both COs and first sergeants were asked who established them. In the majority of the cases they were originated by someone other than the incumbent COs and first sergeants. (3) Then there were cases in which a strong first sergeant assumed responsibility, without deliberate delegation by the CO. (4) Finally, in a few units, the COs disinterest in CI led to the first sergeant assuming responsibility for its gating.

TABLE 4

Question 3 - Who decides?
Correlation of CO vs. First Sergeant
Opinion

Response	Position	
	COs	1st SGs
CO only	31%	37%
XO only	1%	5%
1st SG only	24%	31%
CO and 1st SG	36%	26%
Other	8%	1%

Senior Commander is Gatekeeper

A few units were observed in which a Senior Commander (Battalion or Post-level) had pre-empted the gatekeeping role, for instance by conducting the major formations.

Each of the gating behavior models clearly represented a difference in command style. It would be interesting to know what effect those differences had on unit effectiveness. Pertinent in this regard is the issue of mandatory CI formations.

Mandatory CI?

Some company leaders would like to see the institution of mandatory CI formations. They remember that this was once an Army-wide requirement and they feel that it would strengthen their resolve--or their negotiating position vice other activities--to have a formal, scheduled requirement once again.

Opinions of this recommendation were solicited (Table 5). Only 16% of COs and 30% of first sergeants stated that they would like to see mandatory CI once again. The remainder of those interviewed (84% of the COs and 70% of the first sergeants) replied emphatically "no"--that they were uniquely aware of their troops' information needs and of the constraints affecting the dissemination of that information.

We recommend against any renewed requirement for these reasons:

- Much of the improvement which we observed in the CI program, and in attitudes toward CI among officers and men, seems to result from a free hand and a soft sell.

- The company commander in most Army units is actually assessed more mandatory duty than he or his unit can perform. Furthermore, unit schedules are complex and crowded. Commanders need as much freedom and local authority as they can get. We observed several units with good programs in which CI formations were a seasonal activity; they did not occur at all during training season. These commanders had made the rational decision not to try CI classes for three/four months at a time, but to have good ones when time permitted.

- The CI message has the greatest punch when the commander delivers it--on his own initiative, because he believes the message important. A mandatory formation reduces him to the role of a passive agent, filling a requirement.

A mandatory formation reduces him to the role of a passive agent, filling a requirement.

TABLE 5

Question: Mandatory CI?

Response	COs	1st SGs
Yes	16%	30%
No	84%	70%

Reading Skill

We made no direct measurement of reading skill, but we did ask a series of questions about who can and does read CI publications. The general answer was that second enlistment soldiers read the publications, but that those who had no career commitment did not. At all junior levels there was a serious reading problem. Table 6 applies.

TABLE 6

Question: Who can understand Soldiers?

	Unit Type		
Response	CA	CS	SC
Everybody	97%	39%	97%

This table refers to Soldiers magazine, a publication which is deliberately popularized and has the simplest reading level of all CI publications. Note that, in the opinion of unit leaders, all members can read Soldiers in Combat Arms (CA) and School (SC) units. In Combat Support (CS) units, however, there is a clear recognition that some enlisted men cannot understand Soldiers. Reading all the other

publications studied, there was a clear belief that they were unreadable to a substantial enlisted population.

This suggests that any CI program which depends primarily on printed media will fail to reach a large audience.

Films

We found a surprising lack of interest in films (Table 7).

TABLE 7

Question: Have you ever used any CI films?

Response	Position COs	1st SGS
Yes	45%	52%
No	55%	48%

Only about one half of the units surveyed had ever used films in the CI program. Of those who had used films, many had used them only because they were directed to do so, and fewer than 25% used them as often as once a month. This was true in spite of the fact that films were readily available. An alternative to films might be electronic media.

RECOMMENDATIONS

Electronic Media

It has been widely observed that modern young adults are of a non-reading generation. For whatever original reason, they are much less likely to turn to a printed source for information or entertainment, than was the pre-TV generation. It follows that, if they are to be reached with maximum effectiveness, they will be reached via radio and TV. But the survey found a low level of interest in, or knowledge of existing CI on electronic media. This is partly because post and regional CI programs are faced with a dilemma in using electronic media: they cannot reach their audience until prime-time, and then they are unable to compete with commercial programming. Cost constraints make it unlikely that the Army, or local installations, can compete with network programs for the attention of the soldier audience. Nevertheless, radio is reaching a small segment of the potential audience steadily and effectively, and TV has a massive potential, especially to reach the first enlistment soldier.

Locally-produced TV deserves a special note. It is hard for installations to produce good footage, except for local news and commanders as talking-heads. These last two are of great value.

Finally, the impact of TV and radio on dependents should not be under-estimated. Dependents apparently are more likely to be reached by daytime programming, and may actually find the programming more interesting. They like to know what the Army is doing, who the local leaders are, and what the local units are doing. Soldier husbands may not be very informative in this regard, and families are known to be influential in such vital decisions as reenlistment. We recommend these ideas to the Army.

Feedback

It is a general principle of systems' theory that messages will be effectively transmitted only in the presence of feedback--information which tells whether a message is correctly understood. This is more than a theoretical consideration; perhaps the most effective practical way to improve the performance of human organizations is to systematize feedback. That principle has been widely used to improve

industrial processes, communications networks, and instruction in schools--to name only a few cases.

Several questions in the survey asked whether COs and first sergeants knew what information their troops received and understood. The worst of all possible conditions was found to occur: company leaders thought they knew, but in fact did not know when their communications to troops were received and understood.

We therefore recommend that the Army consider designing into its system a means by which specific key CI messages can be identified, followed through the information system, and their receipt verified by sampling comprehension at the troop level. Feedback studies could be used in normal information operations, or experimentally to determine what information strategies are most effective.

Reinforcement

The hypothesis that COs and first sergeants get little credit for a good CI program was confirmed. Senior commanders often know little about company CI programs--and in some cases do not care. There was little monitoring of unit CI programs in most of the units studied. More significantly, interviewees did not believe that the quality of their CI program was likely to affect their effectiveness reports. We were apparently observing a condition in which unit CI programs were undertaken largely on local initiative (there were outstanding exceptions, in which battalion or more senior commanders vigorously encouraged good CI).

Here again, a feedback device would solve the problem automatically. If a battalion commander really knew in which of his companies the troops were (and were not) getting the word, he would learn a lot both about the CI program and the effectiveness of his subordinate commanders.

POTENTIAL OF THE METHOD

At the beginning we said that the methods of this research should be applied more widely. In fact, we are now applying similar techniques to study one aspect of fire control in the Field Artillery. We hope to see it used particularly to address problems of two kinds: morale, and tactical intelligence.

The applicability to problems of morale is probably clear from the context of this report. But Kinton believes that this method of research might lead to dramatic enhancement of small unit combat potential.

It is a commonplace observation that battle takes place under conditions of inadequate information. Researchers have repeatedly reported what every soldier knows--that each officer and man at the small unit level fights with astonishingly little understanding of the situation and of his role. This is true in spite of the fact that men at the company front may observe information vital to the commander, or to the squad at the flank, and not understand the value of that information nor report it. Staffs at battalion level do not know what information to pass, or when to pass it to the men at company and platoon level.

An observation from REALTRAIN exercises has been that good internal informal communications are one of the features of winning platoons. However, communication, in conventionally trained units, tends to freeze permanently on contact with enemy fire.

This matter has never been given specific study. KINTON recommends it to the Army, and to other researchers.

WORKER-ORIENTED & JOB-ORIENTED INSTRUMENTS FOR EVALUATING JOB PERFORMANCE[1,2]

Robert Vineberg

Elaine N. Taylor

Human Resources Research Organization
(HumRRO)

[1]Paper presented at 1977 Military Testing Association Conference, San Antonio, Texas, 17-21 October 1977.

[2]Sponsored by the Naval Education & Training Command and the Personnel & Training Research Program, Office of Naval Research, Contract N00014-75-C-0938 (NR 156-047). The project monitor was Dr. Marshall Farr, Director, Personnel & Training Research Program.

WORKER-ORIENTED & JOB-ORIENTED INSTRUMENTS FOR EVALUATING JOB PERFORMANCE[1,2]

I am going to describe two instruments for obtaining performance ratings. These two instruments have been developed at the task and element level of jobs. These instruments, the Performance Analysis Inventory (PAI) and the Task Proficiency Inventory (TPI), were developed as part of an ONR study concerned with the performance capabilities of men at different aptitude levels. The part of the study concerning aptitude levels has not yet been undertaken.

In developing these instruments, we collected data on men in ten Navy ratings and three pay grades aboard the aircraft carriers Enterprise and Constellation.

Depending on the job being performed and the type of instrument, the forms contain between 34 and 93 rating items. A particular feature of this study then, is the development of information about performance in many elements of a job, rather than with regard to a few global measures.

The rating instruments are based upon two different models of job analysis. Ernest McCormick (1972) has referred to these approaches as worker-oriented and job-oriented models. A worker-oriented approach focuses on elements of behavior that generalize across tasks and jobs. For example, observing visual displays, obtaining information from written materials, using non-precision tools, activating variable setting controls, following fixed procedures, estimating quantity, analyzing information, or negotiating with people.

A job-oriented approach to job analysis focuses on specific technological elements of job content. For example, repairing carburetors, drafitng business letters, annealing copper tubing, organizing stock control functions, or translating Russian newspaper articles.

Forms of our worker-oriented instrument, the PAI, are based upon elements of jobs taken from McCormick's job analysis questionnaire, the Position Analysis Questionnaire or PAQ. The jobs in the study were first analyzed with a modified form of the PAQ that we developed for use in the Navy. Each job was analyzed by rating the relevance or importance of each of 139 possible worker-oriented elements. Then, performance rating scales were developed for each element of importance that emerged.

---

[1] Paper presented at 1977 Military Testing Association Conference, San Antonio, Texas, 17-21 October 1977.

[2] Sponsored by the Naval Education & Training Command and the Personnel & Training Research Program, Office of Naval Research, Contract N00014-75-C-0938 (NR 156-047). The project monitor was Dr. Marshall Farr, Director, Personnel & Training Research Program.

TABLE 1

This slide shows the jobs in the study and the number of worker-oriented items included in the various forms of the Performance Analysis Inventory for obtaining performance evaluations.

Forms of our job-oriented rating instrument, the Task Proficiency Inventory, were based on task inventory data furnished by the Navy Occupational Task Analysis Program (NOTAP). Here, we used existing task analysis data to identify elements of performance to be evaluated.

OVERLAY TO TABLE 1

This overlay adds the number of job-oriented items included in the various forms of the Task Proficiency Inventory for obtaining performance evaluations. Here, separate instruments were developed by pay grade where such job analysis data was available. At the time we constructed the scales, NOTAP job analysis data were not available for Electrician's Mate, Hull Maintenance Technician, and Interior Communications. The essential feature of both methodologies is that performance is defined and evaluated in terms of very specific behavioral or technological referents.

TABLE 2

This slide shows some sample items from the worker-oriented PAI for Aviation Boatswain's Mate - Equipment.

TABLE 3

This slide shows some sample items from the job-oriented TPI for the same job.

To provide a basis for item analysis, performance data were obtained for 569 incumbents in the ten jobs. For comparative purposes, we also obtained Performance Evaluation Report scores. This is the instrument that is used operationally in the Navy for evaluating a man's performance.

FIGURE 1a

This slide shows frequency distributions of scale value usage for E-3 and E-4 on the different instruments. The worker-oriented PAI and the job-oriented TPI have seven-point scales, while the Performance Evaluation Report has ten-point scales. The distributions are displayed with the mid-points of the scales coinciding in order to avoid the distortion that occurs if data from one type of scale are expressed in terms of the other.

The PAI & TPI show less skew than the operational instrument, but it must be pointed out that here we are comparing experimental and operational data. Obviously, we do not know what characteristics our instruments would demonstrate if they were administered on a continuing basis by military personnel.

FIGURE 1b

This slide shows the distributions for E-5 and all pay grades combined. As always, skewness increases with grade. The job element or task level approach has not overcome this problem. Frequency data in the handout shows respectable distributions for our instruments at the E-3 level, but deterioration setting in at the E-4 level.

The handout also contains tables of means, standard deviations of subject means, subject standard deviations, and item standard deviation. These analyses were undertaken to look for relative leniency, halo, and discrimination between the two experimental rating instruments. Comparisons among these statistics show less leniency and halo and better discrimination for the worker-oriented scales than the job-oriented scales.

This completes my presentation. We presently are awaiting permission to use these instruments in the second phase of our study to collect data on the performance of men at different aptitude levels.

REFERENCE

Ernest J. McCormick, Paul R. Jeanneret, & Robert C. Mecham, "A study of job characteristics and job dimensions as based on the position analysis questionnaire (PAQ)", monograph, *Journal of Applied Psychology*, Vol. 56, No. 4, August 1972.

TABLE 1

NUMBER OF ITEMS IN THE PERFORMANCE ANALYSIS INVENTORY (PAI) & THE TASK PROFICIENCY INVENTORY (TPI) BY NAVY JOB & PAY GRADE.

NAVY JOB	PAI	TPI		
		E3	E4	E5
Aviation Boatswain's Mate				
Equipment (ABE)	56	40	40	63
Fuel (ABF)	49	40	40	50
Handling (ABH)	56	40	40	50
Aviation Ordnance (AO)	45	54	40	51
Electrician's Mate (EM)	50	—	—	—
Hull Maintenance Technician (HT)	47	—	—	—
Interior Communication (IC)	49	—	—	—
Mess Management Specialist				
(MS-S2 Division)	39	54	40	56
(MS-S5 Division)	38	93	68	86
Storekeeper (SK)	34	56	40	61

OVERLAY TO TABLE 1

TABLE 2

SAMPLE ITEMS FROM WORKER-ORIENTED PAI FOR AVIATION BOATSWAIN'S MATE - EQUIPMENT

2. Work produced using energy-powered tools to perform operations not requiring great accuracy or precision. (Electric grinders and drills, welding equipment, brazing gear, skill saw, etc.)

Exceptionally Good			Satisfactory			Exceptionally Poor	Never Has to
7	6	5	4	3	2	1	X

3. Work accomplished using handling devices. (Pouring zinc from ladles, using mechanical fingers, etc.)

Exceptionally Efficient			Satisfactory			Exceptionally Inefficient	Never Has to
7	6	5	4	3	2	1	X

37. Remembering information for a brief period of time. (Console recorder, launch valve strobe timer readings, steam pressures, etc.)

Very Reliable			Satisfactory			Very Unreliable
7	6	5	4	3	2	1

41. Being aware of and alert to the condition/quality of equipment material, or weapon systems. (For example, condition of components in catapult and recovery gear, etc.)

Exceptionally Aware			Satisfactory			Exceptionally Unaware
7	6	5	4	3	2	1

42. Being accurate in transcribing. (Copying or posting data or information for later use; water brake readings, fluid history reports, etc.)

Exceptionally Accurate			Satisfactory			Very Inaccurate	Never Has to
7	6	5	4	3	2	1	X

53. Obtaining job information by seeing differences using far vision. (Deck edge operator, aircraft identification to determine correct settings for arresting gear, etc.)

Exceptionally Good			Satisfactory			Exceptionally Poor	Never Has to
7	6	5	4	3	2	1	X

TABLE 3

SAMPLE ITEMS FROM JOB-ORIENTED TPI FOR
AVIATION BOATSWAIN'S MATE - EQUIPMENT

Very Effective			Average			Very Ineffective	Never Has to
7	6	5	4	3	2	1	X

- ☐ 17. Applying preservatives to cables (CDPS, purchase cables, bridles, etc.).
- ☐ 18. Cleaning hydraulic filters.
- ☐ 19. Stowing/breaking out parts/equipment.
- ☐ 20. Rigging the barricade.
- ☐ 21. Changing Bridle arrestor straps.
- ☐ 22. Replacing "O" rings in valves/cylinders.
- ☐ 23. Painting safety markings on flight deck.
- ☐ 24. Repacking the retract valve.
- ☐ 25. Maintaining logs/records (catapult, flight deck, fuels, etc.).
- ☐ 26. Participating in working parties.
- ☐ 27. Functionally checking catapults by firing no-loads.
- ☐ 28. Ensuring safety lines are in place during no-load firings.
- ☐ 29. Changing purchase cable on "AG" engines (re-reeve).
- ☐ 30. Safety wiring equipment/gear/switches.
- ☐ 31. Measuring slipper wear.

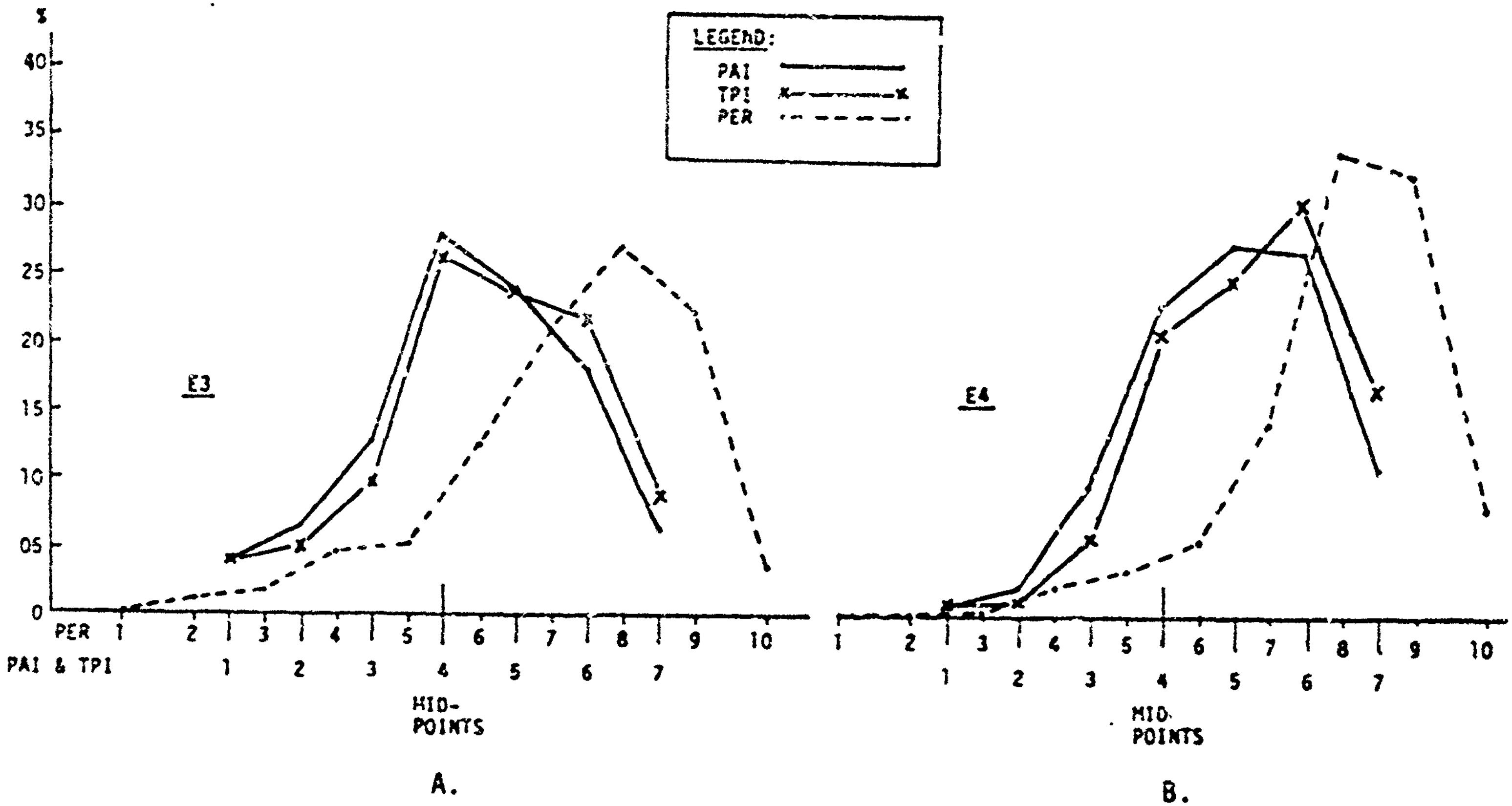


Figure 1a. Frequency Distributions (in %) of Scale Value Usage for Three Rating Instruments Drawn With the Mid-Points of Scales Coinciding.

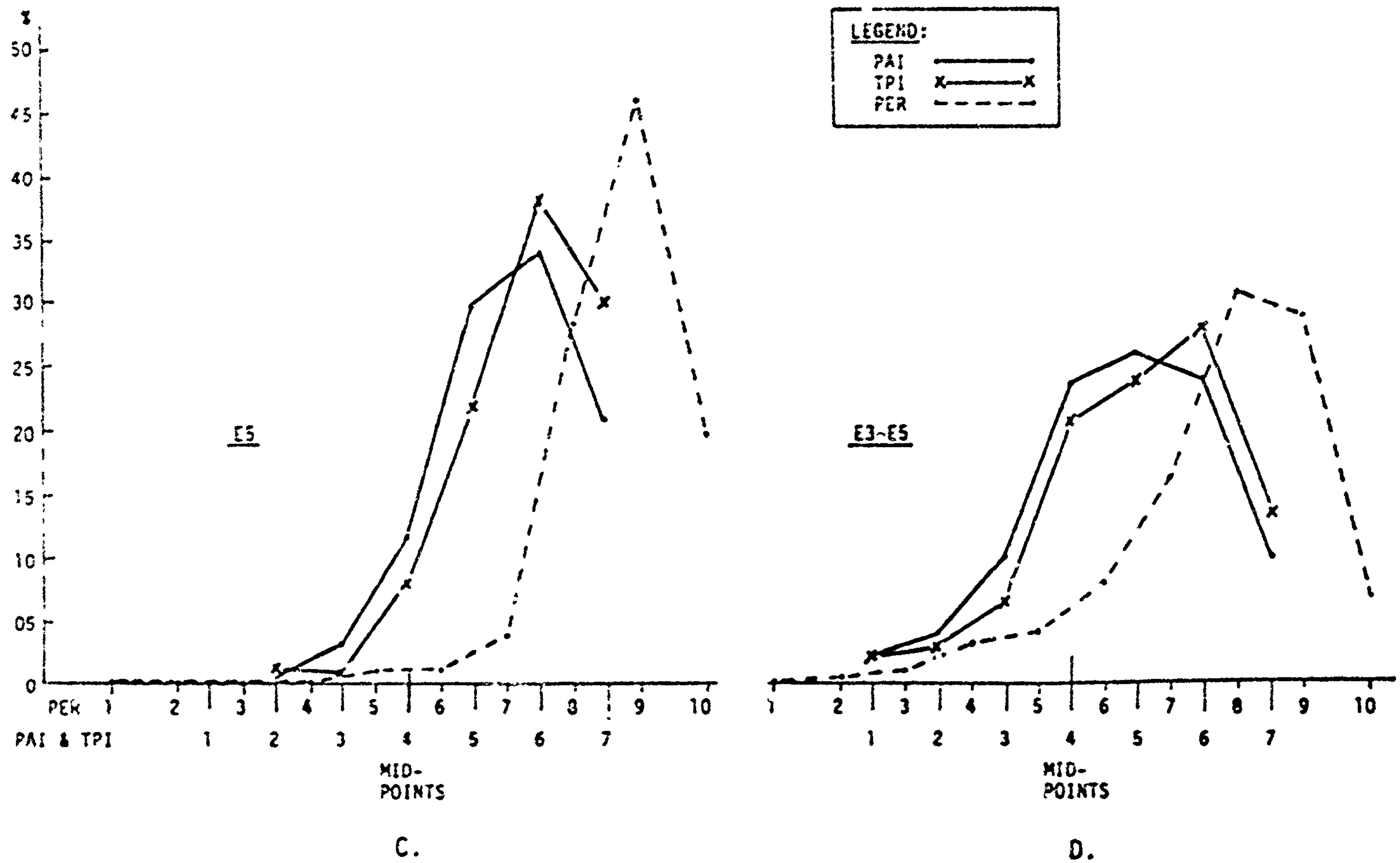


Figure 1b. Frequency Distributions (in %) of Scale Value Usage for Three Rating Instruments Drawn With the Mid-Points of Scales Coinciding.

TABLE 4. FREQUENCY OF USAGE OF SCALE VALUES FROM E3 DATA ON THREE RATING INSTRUMENTS (ALL NAVY JOBS, ALL ITEMS)

	PAI		TPI		PER		
SCALE VALUES	f	%	f	%	NAVY SCALE VALUE	f	%
1	420	4.14	144	4.02	1.0	1	.16
2	686	6.77	180	5.03	2.0	7	1.10
3	1306	12.88	355	9.92	2.6	12	1.88
4	2856	28.18	943	26.35	2.8	31	4.85
5	2394	23.62	851	23.78	3.0	34	5.32
6	1841	18.17	786	21.96	3.2	80	12.52
7	631	6.23	320	8.94	3.4	135	21.13
					3.6	174	27.23
Never Has To	1615		1585		3.8	142	22.22
					4.0	23	3.60
TOTAL*	10134	99.99	3579	100.00		639	100.01

*Totals do not include "Never Has To" perform.

TABLE 5. FREQUENCY OF USAGE OF SCALE VALUES FROM E4 DATA ON THREE RATING INSTRUMENTS (ALL NAVY JOBS, ALL ITEMS)

	PAI		TPI		PER		
SCALE VALUES	f	%	f	%	NAVY SCALE VALUE	f	%
1	95	0.92	37	1.00	1.0	—	—
2	241	2.32	50	1.35	2.0	—	—
3	999	9.63	209	5.63	2.6	4	.46
4	2349	22.65	762	20.53	2.8	18	2.09
5	2821	27.20	909	24.50	3.0	30	3.48
6	2758	26.60	1122	30.23	3.2	47	5.45
7	1107	10.68	622	16.76	3.4	123	14.27
					3.6	292	33.83
Never Has To	683		1444		3.8	279	32.37
					4.0	69	8.01
TOTAL*	10370	100.00	3711	100.00		862	100.01

*Totals do not include "Never Has To" perform.

TABLE 6. FREQUENCY OF USAGE OF SCALE VALUES FROM E5 DATA ON THREE RATING INSTRUMENTS (ALL NAVY JOBS, ALL ITEMS)

	PAI		TPI		PER		
SCALE VALUES	f	%	f	%	NAVY SCALE VALUE	f	%
1	4	0.13	2	0.13	1.0	—	—
2	14	0.46	16	1.07	2.0	—	—
3	98	3.18	13	0.87	2.6	—	—
4	355	11.53	119	7.96	2.8	—	—
5	915	29.71	326	21.79	3.0	1	.98
6	1052	34.16	571	38.17	3.2	1	.98
7	642	20.84	449	30.01	3.4	4	3.92
					3.6	29	28.43
Never Has To	118		459		3.8	47	46.08
					4.0	20	19.61
TOTAL*	3080	100.01	1496	99.99		102	100.00

*Totals do not include "Never Has To" perform.

TABLE 7. FREQUENCY OF USAGE OF SCALE VALUES FROM E3-E5 DATA ON THREE RATING INSTRUMENTS (ALL NAVY JOBS, ALL ITEMS)

	PAI		TPI		PER		
SCALE VALUES	f	%	f	%	NAVY SCALE VALUE	f	%
1	519	2.20	183	2.08	1.0	1	.06
2	941	3.99	246	2.80	2.0	7	.44
3	2403	10.19	577	6.57	2.6	16	1.00
4	5560	23.58	1824	20.76	2.8	49	3.06
5	6130	25.99	2086	23.74	3.0	65	4.06
6	5651	23.96	2479	28.22	3.2	128	7.99
7	2380	10.09	1391	15.83	3.4	262	16.34
					3.6	495	30.88
Never Has To	2416		3488		3.8	468	29.20
					4.0	112	6.99
TOTAL*	23584	100.00	8786	100.00		1603	100.02

*Totals do not include "Never Has To" perform.

TABLE 8. MEANS & STANDARD DEVIATIONS OF SUBJECT MEANS ON THREE RATING INSTRUMENTS.

	MEAN			STANDARD DEVIATION			N		Raw
E3	PAI	TPI	Converted PER	PAI	TPI	PER	PAI & TPI	N PER	PER Mean
ABE	4.44	4.71	5.05	1.29	1.26	.94	46	29	3.42
ABF	4.13	4.35	5.35	1.23	1.13	.68	23	13	3.51
ABH	4.64	4.97	5.22	1.18	1.04	1.06	42	33	3.47
AO	4.30	4.77	5.28	1.52	1.37	.83	37	24	3.49
EM	4.36	—	4.94	1.00	—	1.19	17	13	3.35
HT	4.29	—	5.78	1.28	—	.66	19	9	3.63
IC	4.73	—	5.39	.75	—	.33	8	6	3.52
MS-S2	3.47	3.82	4.86	1.22	1.29	1.20	20	12	3.36
MS-S5	4.29	4.56	5.38	1.33	1.34	.90	33	17	3.52
SK	4.16	4.63	4.77	1.23	1.03	.88	10	5	3.30
E4									
ABE	5.19	5.25	5.51	.89	.92	.73	22	12	3.55
ABF	5.00	5.14	5.67	.87	.92	.64	28	18	3.60
ABH	5.31	5.31	6.01	1.12	1.12	.45	30	24	3.70
AO	4.87	5.02	5.69	1.29	1.27	.63	36	28	3.61
EM	4.40	—	5.26	1.10	—	.67	31	30	3.48
HT	4.76	—	5.63	.88	—	.94	24	22	3.60
IC	4.45	—	5.30	.99	—	.63	15	13	3.49
MS-S2	5.01	5.12	5.73	1.09	1.03	.64	23	16	3.62
MS-S5	5.51	5.90	6.49	.56	.64	.59	12	8	3.85
SK	5.12	5.66	5.97	.87	.72	.53	18	14	3.69
E5									
ABE	5.45	5.58	6.50	.99	.87	NC*	9	3	3.89
ABF	5.61	5.74	5.87	.63	.59	.17	8	4	3.66
ABH	5.78	5.86	**	NC	NC	**	3	**	**
AO	6.20	6.14	**	1.00	1.07	**	5	**	**
EM	5.03	—	6.07	.26	—	.58	10	4	3.72
HT	NC	—	NC	NC	—	NC	2	2	NC
IC	4.67	—	NC	NC	—	NC	3	2	NC
MS-S2	5.35	5.54	NC	.85	.86	NC	11	1	NC
MS-S5	6.00	6.26	**	.87	.60	**	17	**	**
SK	5.32	5.73	6.23	1.19	1.09	.61	7	4	3.77
E3-E5									
ABE	4.77	4.96	5.28	1.22	1.16	.94	77	45	3.49
ABF	4.74	4.91	5.57	1.12	1.08	.64	59	35	3.57
ABH	4.95	5.14	5.55	1.19	1.07	.94	75	57	3.57
AO	4.69	4.97	5.51	1.45	1.34	.75	78	52	3.55
EM	4.49	—	5.22	1.00	—	.87	58	48	3.47
HT	4.62	—	5.72	1.05	—	.86	45	34	3.62
IC	4.56	—	5.35	.87	—	.56	26	22	3.51
MS-S2	4.51	4.72	5.36	1.35	1.30	.98	54	29	3.51
MS-S5	5.00	5.28	5.74	1.34	1.32	.96	62	25	3.62
SK	4.89	5.38	5.78	1.12	.99	.81	35	23	3.63
OVERALL	4.73	5.04	5.49	1.23	1.20	.86	PAI 569 TPI 440	370	

*The number of cases was so low, that the statistic was not computed where NC is shown.
**No data were available.

TABLE 9. NUMBER OF TIMES MEANS ON ONE RATING INSTRUMENT EXCEED ANOTHER

	PAI > TPI	PAI > PER	TPI > PER
E3	0/7	0/10	0/7
E4	0/6*	0/10	0/7
E5	1/7	0/7**	0/4**

*A tie occurred in one comparison.

**PER data were not available for 3 Navy jobs, thus reducing the number of compariosns that could be made between PAI and PER to 7, and between TPI and PER to 4.

TABLE 10. NUMBER OF TIMES STANDARD DEVIATIONS OF MEANS ON ONE RATING INSTRUMENT EXCEED ANOTHER

	PAI > TPI	PAI > PER	TPI > PER
E3	5/7	9/10	6/7
E4	3/6	8/10	7/7
E5	4/6	2/3	2/2

TABLE 11. MEAN OF SUBJECT STANDARD DEVIATIONS FOR THREE RATING INSTRUMENTS

	E3	E4	E5	E3-E5
PAI				
ABE	0.84	0.94	0.59	0.84
ABF	0.72	0.67	0.69	0.69
ABH	0.86	0.74	0.63	0.80
AO	0.65	0.67	0.73	0.67
EM	0.75	0.79	0.68	0.76
HT	0.55	0.64	NC*	0.60
IC	0.71	0.71	0.75	0.71
MS-S2	0.70	0.71	0.62	0.69
MS-S5	0.69	0.68	0.60	0.66
SK	0.76	0.73	0.66	0.73
TPI				
ABE	0.80	0.83	0.77	0.81
ABF	0.76	0.66	0.65	0.70
ABH	0.74	0.78	0.69	0.75
AO	0.61	0.68	0.62	0.64
EM	**	**	**	**
HT	**	**	**	**
IC	**	**	**	**
MS-S2	0.62	0.71	0.56	0.65
MS-S5	0.57	0.61	0.56	0.57
SK	0.51	0.67	0.66	0.62
PER				
ABE	0.47	0.66	0.29	0.51
ABF	0.67	0.48	0.31	0.54
ABH	0.60	0.47	NC	NC
AO	0.50	0.54	NC	NC
EM	0.47	0.60	0.66	0.57
HT	0.36	0.44	NC	NC
IC	0.56	0.52	NC	NC
MS-S2	0.43	0.47	NC	NC
MS-S5	0.33	0.33	NC	NC
SK	0.47	0.35	0.46	0.40

*The number of cases was so low, that the statistic was not computed where NC is shown.

**No data were available.

TABLE 12. NUMBER OF TIMES AVERAGE SUBJECT STANDARD DEVIATIONS ON ONE RATING INSTRUMENT EXCEED ANOTHER FOR A GIVEN PAY GRADE

	PAI > TPI	PAI > PER	TPI > PER
E3	6/7	10/10	7/7
E4	4/6	10/10	7/7
E5	5/7	4/4	3/3

TABLE 13. MEAN OF ITEM STANDARD DEVIATIONS FOR THREE RATING INSTRUMENTS.

	E3	E4	E5	E3-E5
	$\bar{X}_{SD}$	$\bar{X}_{SD}$	$\bar{X}_{SD}$	$\bar{X}_{SD}$
PAI				
ABE	1.52	1.18	1.14	1.42
ABF	1.38	1.09	0.88	1.26
ABH	1.40	1.27	0.62	1.36
AO	1.60	1.42	1.17	1.56
EM	1.26	1.35	0.74	1.21
HT	1.42	1.12	0.16	1.24
IC	1.03	1.15	0.86	1.10
MS-S2	1.36	1.31	1.00	1.49
MS-S5	1.39	0.85	0.98	1.31
SK	1.34	1.07	1.23	1.27
TPI				
ABE	1.45	1.19	1.07	1.28
ABF	1.26	1.15	0.81	1.24
ABH	1.26	1.27	0.49	1.09
AO	1.52	1.44	0.97	1.18
EM	**	**	**	**
HT	**	**	**	**
IC	**	**	**	**
MS-S2	1.52	1.22	0.96	1.45
MS-S5	1.52	0.80	0.84	1.21
SK	1.05	1.00	1.05	0.98
PER				
ABE	1.05	0.88	0.38	1.03
ABF	0.72	0.78	0.40	0.81
ABH	1.07	0.60	NC*	0.98
AO	1.05	0.82	NC	0.94
EM	1.36	0.88	0.75	1.05
HT	0.61	1.02	0.28	0.97
IC	0.55	0.81	0.56	0.75
MS-S2	1.26	0.80	NC	1.03
MS-S5	0.98	0.61	NC	0.92
SK	1.03	0.70	0.78	0.92

*The number of cases was so low, that the statistic was not computed where NC is shown.

**No data were available.

TABLE 14. NUMBER OF TIMES AVERAGE ITEM STANDARD DEVIATIONS ON ONE RATING INSTRUMENT EXCEED ANOTHER FOR A GIVEN PAY GRADE

	PAI > TPI	PAI > PER	TPI > PER
E3	7/9	9/10	7/7
E4	3/6	10/10	7/7
E5	7/7	5/7	4/4

Promotion Evaluation for Inter-Organizational Referral:
A Behavioral Expectation Approach

Charles N. MacLane
U.S. Civil Service Commission

Introduction

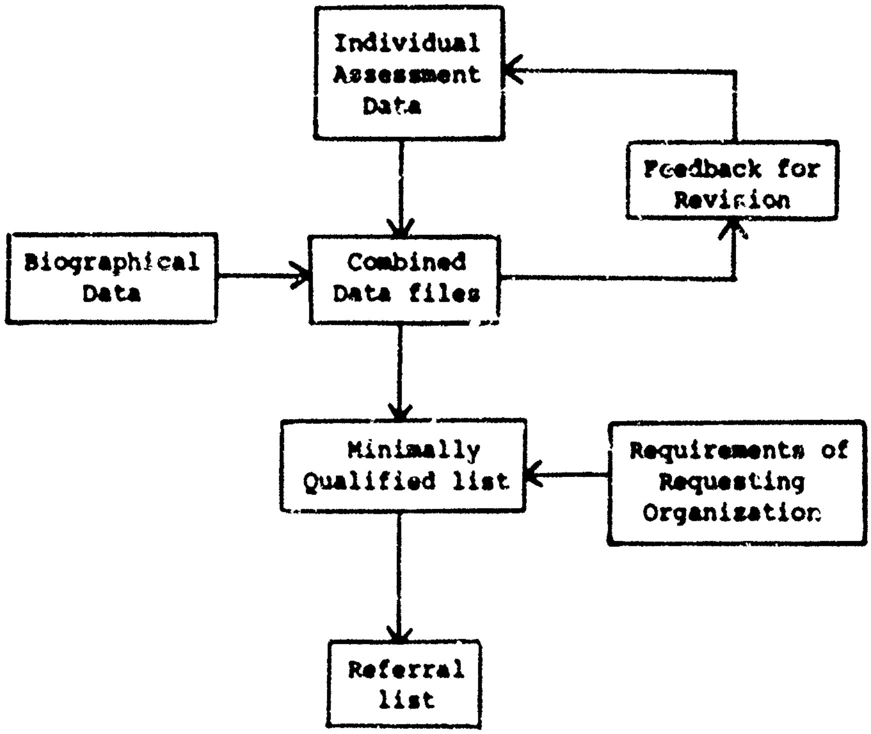
This paper describes the design and the first phase of development of a system for the initial evaluation of Federal personnel specialists for promotion and transfer across organizations. It was created to be part of an existing automated personnel record system (called the Federal Automated Career System) which presently screens individuals on background and experiential data for referral to agencies other than their own. The existing FACS referral process is limited in that it can validly screen individuals for particular positions only in terms of broad, general categories of experience and background. A method of meaningfully assessing individuals within these broad categories was needed.

It was considered important to tailor the methodology to the parameters of the proposed assessment system, especially in view of trends in current research on performance appraisals. Grey and Kipnis (1976), for example, found that variance in supervisory ratings of performance showed a contrast effect; that is, the more employees a supervisor had whom he labeled "bad", the higher the overall mean rating he gave his employees. Thus, unless there is some means of standardizing ratings across organizations, different ratings may be given based on the number of poor employees a supervisor has. Thus it is well to consider carefully the possibly differential effects of contextual variables.

In the present study, the extent of coverage of the proposed system (about 12,000 personnel specialists in four major specialty areas and located in hundreds of different organizational settings) was seen as the primary variable with which the methodology had to deal. This in turn raised two issues: 1) Most important was to establish that there were dimensions of personnel work and of things which workers did within these dimensions which were common to all or many of the organizations involved; and 2) It was necessary to have a method of standardizing performance levels within these dimensions which could be used across organizations. If these issues could be resolved, it would provide a basis for insuring the validity of the appraisal process. However, given the extent of coverage, it was not expected that a single study would provide a complete enough sampling of the behavioral domain to completely address the two issues raised above. Therefore, a multi-stage developmental process was planned in which the initial operation of the system would provide data needed for subsequent stages.

Figure I

Algorithm of Assessment Process



In general, then, the project would proceed as follows:

1. A first-stage assessment instrument would be developed which would provide ratings on dimensions of personnel work common to the organizations involved.

2. There would also need to be a mechanism for collecting further data on the work dimensions which would be collected simultaneously with the assessments.

3. Data would be stored on each individual in terms of each dimension and this would interface with biographical data.

4. Operationally, a request for referrals would be made in terms of dimensions critical to the open position and a referral list generated based on the sum of critical dimension scores plus the biographical data.

5. After some period of operational time, the data on the work dimension would be used to revise and expand the instrument. This process would be repeated whenever it was felt necessary. Figure 1. illustrates a generalized model of the operational system.

Methodology

It was decided to adopt the Behavioral Expectation Methodology of Smith and Kendall (1963) because the development and operational use of this type of performance appraisal would provide much of the data required.

For the present study, a sample of 69 personnel specialists (representative of the grade levels, specialities and organizations which were to be included in the system) wrote over 500 short, critical-incident-type statements describing highly effective, moderately effective, and ineffective behaviors in personnel work. In addition, 23 broad dimensions or factors of personnel work were developed by a panel of five very knowledgeable personnel specialists from lists of dimensions developed previously. (In other studies utilizing this methodology, the dimensions have been derived from groupings of the behavioral statements themselves.) The 69 specialists then placed each behavioral statement under the one of the 23 dimensions which it best exemplified and, subsequently, indicated on a 1-7 point scale, the level of effective performance which that statement represented in terms of the dimension into which it had been placed.

Table 1

Personnel Work Dimensions and Means and Standard Deviations of Associated Behavioral Statements

Dimensions	Behavioral Statements					
	Ineffective		Moderately Effective		Highly Effective	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
I. RECRUITMENT NEEDS (Analyzing staffing patterns, anticipating vacancies, maintaining applicant supply, and developing recruitment sources)	2.4	1.1	3.75	1.0	5.7	1.6
II. POSITION-PERSON MATCHING (Matching candidates' qualifications with job requirements and organizational needs)	1.7	.8	4.9	1.4	6.7	.5
III. INFORMATION EXCHANGE (Providing information to employees and the public on personnel procedures, practices, rules and regulations)	2.1	1.1	4.6	1.2	6.3	.7
IV. ADVISORY SERVICES TO MANAGEMENT (Assisting managers in carrying out their responsibilities for personnel management)	2.8	1.2	4.8	.9	5.8	1.0
V. JOB DESIGN AND REDESIGN (Position Management)	1.6	.8	4.6	.8	5.6	1.1
VI. EMPLOYEE PROBLEM-SOLVING	2.8	.8	4.2	1.0	5.8	.8

Table 1 (cont.)

Personnel Work Dimensions and Means and Standard Deviations of Associated Behavioral Statements

Dimensions	Behavioral Statements					
	Ineffective		Moderately Effective		Highly Effective	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
VII. TRAINING DELIVERY (Planning, conducting and evaluating training)	2.1	.9	4.6	1.0	5.9	.9
VIII. TITLE, SERIES, AND GRADE DETERMINATION (Job analysis and evaluation)	1.7	.7	4.3	.9	5.7	1.1
IX. NEGOTIATING AND DEALING WITH UNIONS	3.0	1.0	4.7	1.1	6.5	.6
X. SPECIAL EMPHASIS PROGRAM (handicapped, upward mobility, V.E.V., etc.)	1.2	1.1	4.7	.9	6.1	.8
XI. IDENTIFYING NEEDS AND OBTAINING DEVELOPMENTAL ACTIVITIES TO MEET THEM	2.7	[illegible]	4.9	.9	6.4	.6

The first type of data obtained through these procedures was the number of placements of each behavioral statement under each dimension as a proportion of the total number of placements of that statement. Statements and dimensions, were retained, altered, or discarded based on a 60% criterion through what might be termed a qualitative cluster analysis. The result was that 11 dimensions (each exemplified by nine to thirteen statements) remained. Thus, it had been objectively established that personnel work in the agencies represented in this sample of 69 personnel specialists could be, at least in part, described by these 11 dimensions and that the dimensions had similar meanings in common, in terms of behavior, across these organizations. Further, the qualitative cluster analysis provided one check on the validity of both the dimensions and behavioral statements.

The second type of data gathered - means and standard deviations of the effectiveness ratings on the seven-point scale-addressed the issue of standardization of performance levels across organizations. The criteria for retention were that the means of the statement scale scores be distributed so as to represent the whole range of the scale and that the standard deviation be below 1.20. (Only the statements passing the 60% criterion were used.)

Three behavioral statements were able to be retained for each dimension although it had been hoped that five could be used. As indicated earlier, it was expected that there would be difficulty in finding qualified behavioral statements without more sampling of the behavioral domain than was possible in this initial study. Table 1. shows the final dimension definitions and the means and standard deviations of the effectiveness ratings for their statements.

Performance Appraisal Pre-Test

Although scales which had been developed were less inclusive than had been hoped, it was decided to proceed with the pre-test because, given that the selection ratio would be very low, that the instrument was for short-term use, and that there were no assessment procedures available at present, an instrument of low validity could still be of considerable utility.

The first instrument tested was a supervisory rating form based on the retained dimensions and behavioral statements. The primary analysis was of intra-observer reliability. Inter-observer reliability was considered less satisfactory because of problems associated with obtaining raters and recent studies showing that ratings may be affected

Table 2

Comparative Error Statistics - Supervisor and Self-Assessment Data

Dimension	Supervisor	Self-Assessment
	Scalability Coefficients	
I	.55	.55
II	.45	.64
III	.57	.68
IV	.45	.51
V	.46	.46
VI	.60	.56
VII	.31	.62
VIII	.60	.68
IX	.36	.70
X	.18	.42
XI	.51	.62
	Estimated Total Rater Error (%)	
	26.7	8.9

by organizational level, etc. (e.g. Zedeck and Baker, 1972). The Liek and Matthews (1968) developmental scaling procedure was adopted. Since the three behavioral statements examplifying each dimension had been ordered from ineffective to effective, the ratees could be rated on each separately. The Liek and Matthews procedure gives a measure of the inconsistency between the order of rating and the order in which the statements had been scaled. This coefficient of scalability is similar to a Guttman coefficient but it is a finer measure in the sense that it reflects not only the frequency of reversals in the hypothesized "correct" order but it also reflects the severity of particular reversals through differential weighting of each type of ordering. A representative sample of fifty-two raters appraised one-hundred forty seven ratees on each of the three behavioral statements for each of the 11 dimensions. The statements were randomly ordered under their dimension. All the subjects were asked to base their ratings on ratee behaviors and half of them were asked to record these.

The critical findings relating to supervisory ratings were: 1) As Table 2 indicates, several of the scalability coefficients were unacceptably low; 2) Of the total number of ratings made, 26.7% were reversals (i.e. errors) which could be attributed to the rater (as opposed to scale error); and 3) The behavioral statements given were unacceptable. The rater error statistic includes reversals which could reasonably be attributed to rater inattention rather than an incorrect ordering of the behavioral statements. Anecdotal evidence indicated that the supervisor sample had neither the time nor the complete knowledge of subordinate behaviors to enable them to supply behavioral statements and that lack of time and motivation had led to much of the rater error. Thus it was decided that a self-assessment approach using essentially the same format could reduce the type of error found in the supervisory ratings because the raters would be more knowledgeable of their own behavior and motivated to seek new positions for themselves.

The Self-Assessment Instrument

A sample of 62 personnel specialists satisfactorily completed the self-assessment form which was essentially unchanged from the supervisory format excepting the instructions (Figure 2). Comparative data on the supervisory and self-assessment ratings are given in Table 2. The scalabilitycoefficients shown represent the proportional difference between randomly ordered ratings and ratings which follow perfectly the hypothesized continuum of effectiveness. Since all the scalability coefficients except two represent significant improvements over chance, it is more meaningful to look at the trend across the dimensions from the

Figure 2

Self-Assessment Form

BEHAVIORAL STATEMENTS	Inadequate Evidence	(Put an "X" in Boxes to Answer) The typical activity you have written represents a(n)... Higher level of knowledges skills, and interest than the statement at left	equivalent level of knowledges skills, and interest with the statement at left	lower level of knowledges, skills, and interest than the statement at left
SET H - TITLE, SERIES, AND GRADE DETERMINATION (Job analysis and evaluation)				
22. The personnel specialist could be expected to write evaluation statement when classifying supervisory position to justify the basis of the supervisory grade level.				
23. In proposed job description, personnel specialist could be expected to accept supervisor's description of subordinate workforce as "large" without seeking further definition of the term.				
24. The personnel specialist could be expected to select appropriate classification standards for cross referencing and review comparable positions for alignment in evaluating a position in an occupation for which there are no published grading criteria.				

STEP 2. Write one or more of your activities which are typical of your work in this area. (Use the sheets headed "EXAMPLES" pp. 10-13). Write these as case studies which illustrate how you typically operate in this area; that is, your typical level of knowledges, skills, and interest. Include the following information:

1. What the problem or objective was.
2. What you did and approximately how many months ago you did it.
3. The name of someone who can verify this activity and its typicality and how he can be reached (telephone number if possible). He may be a peer, subordinate, supervisor or someone outside the organization.

supervisory to the self-assessment scales. It can be seen that there is improvement in every case but one and in some cases the improvement is considerable. The behavioral statements were significantly improved both in quantity and depth of coverage.

The most important finding was that the criteria used in developing the scales were related to the level of the scalability coefficients found for the self-assessment but not for the supervisory assessments. That is, in constructing the assessment scales, attention was paid to selection of the behavioral statements with minimum standard deviations and means equally spaced along the scale of one to seven.

It is clear from a comparison of the Table 1 means and standard deviations that the scale construction was reasonably successful except in maintaining the appropriate distance between the means of the behavioral statements representing the highly-effective and moderately effective points on the scales for the eleven dimensions. The closer together the means were placed, the more likely it would be that these two levels would be confused and that a reversal would be made in the rating. It was therefore hypothesized that the scalability coefficients (which reflect errors) would be related to the distance between these two means, if the scale were the cause of the errors, and that the scalability coefficients would be unrelated to the other scale construction criteria. Table 3 shows that this prediction is borne out for the self-assessment data and not for the supervisory data. This suggests that the error in the self-assessment scales can be decreased through improvements in construction in the scale where the supervisory instrument could not.

Second Phase

It is now planned to put the self-assessment instrument into operational use for one year. During this time, thousands of personnel specialists will be filling out self-assessment forms and at the same time contributing behavioral statements for further development of the instruments. These behavioral statements will provide a comprehensive sampling of the behavioral domain. The categorization and scaling process can then be repeated. It is expected that the behavioral statements derived from this large sample will allow more complete definitions of the present eleven dimensions and that more dimensions can be developed. At the end of an operational year, the self-assessment form will be revised.

Table 3

Correlational Data Relating Scale Construction Criteria and Coefficients of Scalability

Scale Construction Criteria	Form	
	Supervisory	Self-Assessment
1. Sum of scale standard deviations	.24	-.13
2. High anchor scale minus low anchor scale mean	-.44	-.07
3. Low anchor scale mean	-.33	.35
4. Moderate anchor scale mean	-.36	-.09
5. High anchor scale mean	-.23	.49
6. Moderate anchor scale mean minus low anchor scale mean	-.46	-.37
7. High anchor scale mean minus moderate anchor scale mean[1]	-.01	.63*

*p = .05

[1]It was hypothesized that the majority of variance in errors across the dimensions would be related to this measure and that it would therefore correlate significantly with the scalability coefficients.

Discussion

This project was referred to as multi-phase because a continuing feedback and revision process is envisioned in which the behavioral statements which are generated as part of the assessment process can be analyzed as reflections of the nature of personnel work in the Federal government and, if changes are seen in the type of behavioral statements being received, the instruments can be revised accordingly. The dimensionality of personnel work can be examined across-and-within organizations by factor-analytic or multi-dimensional scaling procedures.

The data which will be collected has implications beyond its use for the assessment system. As Blood (1974) has discussed, the data can be made available for training purposes, for organizational diagnoses, (e.g. for ascertaining how different organizational levels view the work content), and for job analysis. The real value of the whole process must lie in the on-going nature of the assessment procedures, in the built-in feedback and, consequentially, in our increased knowledge of the performance appraisal process as one of many inter-related organizational systems which continuously affect, and are affected by, one another.

Bibliography

Bernardin, H.J. Behavioral Expectation scales versus summated scales; a fairer comparison. Journal of Applied Psychology, 1977, 62, 422-427.

Blood, M.R. Spin-offs from Behavioral Expectation Scale procedures. Journal of Applied Psychology, 1974, 49, 159-170.

Campbell, J.P., Dunnette, M.D., Arvey, R.C. and Helleruik, L.N. The development and evaluation of behaviorally-based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

Grey, R.J., and Kipnis, D. The influence of perceived organizational characteristics on evaluative processes. Journal of Applied Psychology, 1976, 61, 329-335.

Liek, R.R., and Matthews, M. A scale for Developmental Processes in G.M. Maranell (Ed.), Scaling: a sourcebook for behavioral scientists. Chicago: Aldine, 1974.

Smith, P.C. and Kendall, L.M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Zedeck, S.I. and Baker, H.T. Nursing performance as measured by Behavioral Expectation Scales: A multitrait-Multirater Analysis. Organizational Behavior and Human Performance, 1972, 7, 457-466.

EFFICACY OF CERTAIN MEASURES IN PREDICTING ARMY OFFICER PERFORMANCE

Arthur C. F. Gilbert, Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

The leadership research program in the Army Research Institute for the Behavioral and Social Sciences (ARI) has included work in two areas related to this paper. One of these areas has involved the identification of leadership styles (Helme, Willemin, & Grafton, 1971) while the other area involves research on the validation of associate ratings (Parrish & Drucker, 1957, Haggerty, 1963; Gordon & Medland, 1965; Downey, Medland & Yates, 1976). These two broad research domains form the basis for the design of this research.

The objective of this research was to evaluate the efficacy of certain measures obtained in the Officer Basic Course in predicting subsequent on-the-job first duty tour performance with particular emphasis on the value of final course peer ratings in the prediction scheme as compared with the other predictor variables.

The predictor variables were the seven sub-scales of the Officer Evaluation Battery (OEB), peer ratings obtained at mid course, final course peer ratings, and the final course grade obtained in OBC. The seven scales of the OEB are Combat Leadership (Cognitive), Technical-Managerial Leadership (Cognitive), Career Potential (Cognitive), Combat Leadership (Non-Cognitive), Technical Managerial Leadership (Non-Cognitive), Career Potential (Non-Cognitive), and Career Intent. On-the-job performance measures consisted of a special purpose Performance Evaluation Form and a weighted average of Officer Efficiency Report (OER) ratings.

---

[1]The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

PROCEDURE

Data Collection

All officers in the 13 Career Branches who attended the Officer Basic Course in Fiscal Year 1974 were administered the Officer Evaluation Battery. Peer ratings were obtained at the mid-point of the OBC and again at the end of the course. Final course grades were obtained from each OBC in either actual grades or in class standing within each OBC class or both.

The Performance Evaluation Form has been described in detail elsewhere (Gilbert, 1975; Gilbert and Grafton, 1976; Gilbert, Hooper, & Hicks, 1977). Essentially this instrument was designed to yield a measure of overall duty performance and rankings and ratings of potential performance along a number of leadership dimensions. Five of these leadership dimensions correspond to factors derived by Helme, Willemin, and Grafton (1971) and the factors of consideration and initiation of structure identified by Fleishman (1974) and Stogdill (1974). In addition, the form required ratings along the two more global dimensions of combat leadership and technical-managerial leadership identified by the Helme, Willemin, and Grafton research. In Figure 1, the dimensions assessed by the Performance Evaluation Form are shown with the corresponding scale of the Performance Evaluation Form, and the abbreviated title of each scale. A seven point scale adapted from Willemin (1965) shown in Figure 2 was used for each rating. Raters were required to rank seven of the scales in terms of this officer's potential for future performance and then provide ratings in these areas. Three of the scales, duty performance, combat leadership, and technical-managerial leadership required ratings only.

Ratings on the Performance Evaluation Form were obtained from four raters as far as possible. Ratings were requested from the officer's immediate supervision, from a superior officer other than the officer's immediate supervisor but not necessarily the OER indorsing official, and from each of two close associates.

Data Preparation

The Officer Basic Course grades and class standings were equated by ranking the grades of those officer's for whom only class grades were available within the OBC class of which he was a member. These rankings were then converted to standard scores. Where rankings were available they were converted to standard scores within the different OBC classes. Scores were standardized with a mean of 100 and a standard deviation of 20.

PROCEDURE

Data Collection

All officers in the 13 Career Branches who attended the Officer Basic Course in Fiscal Year 1974 were administered the Officer Evaluation Battery. Peer ratings were obtained at the mid-point of the OBC and again at the end of the course. Final course grades were obtained from each OBC in either actual grades or in class standing within each OBC class or both.

The Performance Evaluation Form has been described in detail elsewhere (Gilbert, 1975; Gilbert and Grafton, 1976; Gilbert, Hooper, & Hicks, 1977). Essentially this instrument was designed to yield a measure of overall duty performance and rankings and ratings of potential performance along a number of leadership dimensions. Five of these leadership dimensions correspond to factors derived by Helme, Willemin, and Grafton (1971) and the factors of consideration and initiation of structure identified by Fleishman (1974) and Stogdill (1974). In addition, the form required ratings along the two more global dimensions of combat leadership and technical-managerial leadership identified by the Helme, Willemin, and Grafton research. In Figure 1, the dimensions assessed by the Performance Evaluation Form are shown with the corresponding scale of the Performance Evaluation Form, and the abbreviated title of each scale. A seven point scale adapted from Willemin (1965) shown in Figure 2 was used for each rating. Raters were required to rank seven of the scales in terms of this officer's potential for future performance and then provide ratings in there areas. Three of the scales, duty performance, combat leadership, and technical-managerial leadership required ratings only.

Ratings on the Performance Evaluation Form were obtained from four raters as far as possible. Ratings were requested from the officer's immediate supervision, from a superior officer other than the officer's immediate supervisor but not necessarily the OER indorsing official, and from each of two close associates.

Data Preparation

The Officer Basic Course grades and class standings were equated by ranking the grades of those officer's for whom only class grades were available within the OBC class of which he was a member. These rankings were then converted to standard scores. Where rankings were available they were converted to standard scores within the different OBC classes. Scores were standardized with a mean of 100 and a standard deviation of 20.

Figure 2

OFFICER PERFORMANCE SCALE[1]

Scale Value	Description
7.	OUTSTANDING: Far above the requirements of the situation, suggesting the highest kind of formal recognition through meritorious award, or decoration.
6.	SUPERIOR: Markedly above the requirements of the situation, suggesting formal recognition through a special (favorable) efficiency report, or letter of commendation.
5.	ABOVE AVERAGE: Somewhat above the requirments of the situation, suggesting informal recognition through specific favorable comment in his regular efficiency report, and through informal appreciation or commendation.
4.	AVERAGE: Fully up to the requirements of the situation, suggesting general appreciation (Perhaps mostly unexpressed).
3.	BELOW AVERAGE: Somewhat below the requirements of the situation, though suggesting only the mildest kind of corrective action through informal recommendation for improvement, or through change of duty assignment within the organization.
2.	MARGINAL: Markedly below the requirements of the situation, suggesting formal corrective action through a special (unfavorable) efficiency report, administrative admonition, letter or reprimand, summary court, or transfer out of the organization.
1.	UNSATISFACTORY: Far below the requirements of the situation, suggesting the most drastic kind of formal corrective action through reclassification, demotion, general court, or boarding out of the Army.

[1] Adapted from Willemin (1965)

Completed Performance Evaluation Forms were edited to insure compatability of rankings and ratings. The ranking of an officer's area of greatest potential performance should have had a ranking of "1" and his weakest area should have had a ranking of "7". The highest rating should be given to his strongest area and a lower or equivalent rating should be assigned to his next strongest area and so forth. Here, it should be pointed out that the rating system permitted the rater to give the same rating on the scales of the form. Where discrepancies were encountered between the ratings and rankings, the ratings or the rankings were corrected.

RESULTS AND DISCUSSION

Reliability estimates for each of the ten scales of the Performance Evaluation Form are shown in Table 1. These estimates are based on cases for which all four ratings were available in the 13 Career Branches. Estimates were obtained by averaging the six possible correlations among the four raters and adjusting the resulting average by the Spearman-Brown Prophecy Formula.

The reliability estimates ranged from .70 for the Combat Leadership Scale to .55 for the Logistical Knowledge Scale. These estimates were reported by Gilbert (1977) and support the findings of Willemin (1965).

The correlations between the ten predictor variables and each of the ten scales of the Performance Evaluation are shown in Table 2 for the entire sample. Also, the multiple correlations between the ten predictor variables are shown in the same table. Examination of this table reveals that final course peer ratings yielded the highest zero order correlations with eight of the scales the two exceptions being the Technical - Managerial scale and the Tactical Knowledge scale. Final course peer ratings were most predictive of overall duty performance, combat leadership, and making decisions.

In Table 3, the correlations among the same set of variables and corresponding multiple correlations are shown for the Combat Arms Branches. These branches are Air Defense, Armor, Field Artillery, and Infantry. In this analysis, final course peer ratings yielded higher or equal zero order correlations with the criteria as did other predictors in all but two instances these being the Technical-Managerial Scale and the Logistical Knowledge scale.

The correlations between each set of predictors and each of the ten criteria and corresponding multiple correlations are shown in Table 4 for the branches other than the combat arms branches. Here for all of the criteria, with the exception of the Technical-Managerial scale, the zero order correlations between final course peer ratings and the criteria were higher or equal to the zero order correlations between each of the other variables and each of the criteria.

Table 1

Reliability Estimates for Each
Scale of the Performance Evaluation Form

Scale	Reliability Estimate
Duty Performance	.67
Combat Leadership	.70
Technical-Managerial Leadership	.58
Tactical Knowledge	.68
Understanding Mission	.59
Making Decisions	.66
Defining Functional Roles	.58
Planning and Organizing	.57
Motivating Troops	.60
Logistical Knowledge	.55

Table 2

Correlations Between Each Predictor and Each Scale of the Performance Evaluation Form for the Total Sample
(N = 2,108)

	Predictor[1]										Multiple
	CLC	TMC	CPC	CLNC	TMNC	CPNC	CI	OBCG	PRM	PRF	Correlation R
Duty Performance	.10**	.04*	.07**	.10**	.08**	.04*	.01	19**	.16**	.26**	29**
Combat Leadership	.21**	.07**	.11**	.26**	.15**	22**	06**	1[illegible]**	18**	.25**	.39**
Technical-Managerial Leadership	.08**	.08**	.08**	.02	07**	-04*	-05*	19**	.10**	.18**	.25**
Tactical Knowledge	.23**	10**	12**	27**	15**	22**	09**	.19**	.14**	.22**	.38**
Understanding Mission	09**	.01	05*	.09**	.07**	.04*	.00	.19**	.15**	.21**	.26**
Making Decisions	.11**	.01	.05**	.14**	12**	.08**	.01	.18**	16**	.26**	.31**
Defining Roles	05**	02	.03	.04*	06**	-.01	-.02	.14**	.12**	.18**	.21**
Planning and Organizing	.07*	.02	.05*	.04*	05*	-.01	-.03	.17**	.12**	.21**	.25**
Motivating Troops	.03	-06**	-04*	.06**	.05*	.03	.01	.10**	.18**	.25**	.28**
Logistical Knowledge	.12**	.09**	11**	05**	08**	-.01	-.04*	.16**	.08**	.18**	.24**

[1]Predictor CLC = Combat Lead. (Cog)
TMC = Tech-Manag. Lead. (Cog)
CPC = Career Pot. (Cog)
CLNC = Tech-Manag. Lead. (Non-Cog)
TMNC = Tech-Manag. Lead. (Non-Cog)
CPNC = Career Pot (Non-Cog)
CI = Career Intent
OBCG = OBC Course Grade
PRM = Mid Career Peer Ratings
PRF = Final Peer Ratings

*Significant at the .05 level
**Significant at the .01 level

Table 7

Correlation Between Each Predictor and Each Scale of the Performance Evaluation Form and Corresponding Multiple Correlation for the Combat Arms Branches
(N = 1,193)

Criterion	Predictor [1]										Multiple Correlation R
	CLC	TMC	CPC	CLNC	TMNC	CPNC	CI	OBCG	PRM	PRF	
Duty Performance	.08**	.03	.06*	.05*	.05*	.02	.00	.21**	.13**	.23**	.27**
Combat Leadership	.12**	.03	.07*	.13**	.10**	.11**	.03	.21**	.16**	.26**	.31**
Technical-Managerial Leadership	.06**	.08**	.05*	-.02	03	-.03*	-.05*	.17**	.09**	.14**	.22**
Tactical Knowledge	.14**	.08**	.10**	.12**	.11**	.13**	.05*	.23**	.13**	.25**	.31**
Understanding Mission	.05*	.02	.04	.05	.05*	.01	-.02	.19**	.11**	.19**	.24**
Making Decisions	.06*	.01	.03	.07**	.08**	.05	-.01	.20**	.13**	.25**	.29**
Defining Roles	.03	.03	.04	-.00	.04	-.03	-.03	.15**	.09	.16**	.21**
Planning and Organizing	.07**	.04	.05*	.00	.02	-.02	.17**	.08**	.08**	.17**	.23**
Motivating Troops	.02	-.03	-.03	.03	.05	.01	-.02	14**	16**	23**	.26**
Logistical Knowledge	.08**	.08**	.09**	.02	.04	-.02	-.02	15**	.02	.11**	.19**

[1] Predictor CLC = Combat Lead. (Cog)
TMC = Tech-Manag. Lead. (Cog)
CPC = Career Pot. (Cog)
CLNC=Tech-Manag. Lead. (Non-Cog)
TMNC=Tech-Manag. Lead. (Non-Cog)
CPNC = Career Pot (Non-Cog)
CI = Career Intent
OBCG = OBC Course Grade
PRM = Mid Career Peer Ratings
PRF = Final Peer Ratings

*Significant at the .05 level
**Significant at the .01 level

Table 4

Correlation Between Each Predictor and Each Scale of the Performance Evaluation Form for Branches Other Than Combat Arms
(N = 915)

Variables	Predictor[1]										Multiple Correlation R
	CLC	TMC	CPC	CLNC	TMNC	CPNC	CI	OBCG	PRM	PRF	
Duty Performance	.11**	.04	.07	.12**	.11**	.03	.01	.16**	.20**	.30**	.33
Combat Leadership	.22**	.07	.15**	.31**	.20**	.18**	.05	.17**	.20**	.31**	.44
Technical-Managerial Leadership	.13**	.08*	.11**	07	.12**	-.01	-.04	.23**	.12**	.23**	.30
Tactical Knowledge	.23**	.08*	.14**	.31**	.20**	.17**	.07	.13**	.17**	.26**	.41
Understanding Mission	.10*	-.01	.05	.10*	.09*	.04	.01	.18**	.19**	.25**	.30
Making Decisions	.12**	.00	.07	.17**	.16**	.05	.02	.15**	.20**	.29**	.35
Defining Roles	.07	-.00	.03	.08*	.09*	.00	-.01	.13**	.15**	.21**	.24
Planning and Organizing	.08*	.00	.04	.08*	.08*	.01	-.01	.15**	.16**	.26**	.29
Motivating Troops	.02	-.10*	-.05	.07	.05	.02	.04	.04	21**	29**	.32
Logistical Knowledge	.17**	.10*	14**	.11**	.12**	.00	-.05	.17**	.14**	.25**	.32

[1]Predictor CLC = Combat Lead. (Cog)
TMC = Tech-Manag. Lead. (Cog)
CPC = Career Pot. (Cog)
CLNC = Tech-Manag. Lead. (Non-Cog)
TMNC = Tech-Manag. Lead. (Non-Cog)
CPNC = Career Pot (Non-Cog)
CI = Career Intent
OBCG = OBC Course Grade
PRM = Mid Career Peer Ratings
PRF = Final Peer Ratings

*Significant at the .05 level
**Significant at the .01 level

In general, the results of these analyses indicate that final course peer ratings obtained in the Officer Basic Course are the best predictor of duty performance when measured by the duty performance scale of the Performance Evaluation Form for the total sample. The zero-order correlation in this instance is .26 between final course peer ratings and duty performance while the next highest zero-order correlation is .19 between the same criterion and OBC final grades. In the Combat Arms Branches there is little difference between the predictive value of peer ratings and OBC final grades when overall duty performance is concerned. However, for branches other than the Combat Arms the correlation between final course ratings and overall performance is .30 while that of grades and the criterion is .16.

The last analyses involved the relationships between the predictors and the weighted average Officer Efficiency Report (OER) ratings. The results of these analyses are shown in Table 5. For the total sample final course peer ratings yielded a correlation of .21 with OER ratings but the final grades obtained in the Officer Basic Course (OBC) yield a correlation of .20. In the Combat Arms Branches OBC final course grades yielded a slightly higher zero order correlation with OER ratings while in the branches other than Combat Arms the reverse obtained.

The results of this research are similar to that reported elsewhere on the utility of associate ratings or peer ratings in predicting subsequent performance (Helme, 1965; Gilbert, 1975; Gilbert and Downey, 1977). Further research will seek to explore how the predictive utility of peer ratings may be enhanced. Possibilities in this regard is to divide the sample into individual career branches since some differences in predictive power between Combat Arms branches and the other branches were observed in this research. Another possibility along this line is to divide the sample according to the similarities of specialties in which the officers are engaged.

Table 5

Correlation Between Each Predictor and Weighted Average Officer Efficiency Ratings and Corresponding Multiple Correlations for the Combat Arms, Branches Other Than Combat Arms, and for the Total Sample

Predictor	Combat Arms (N =2,486)	Branches Other Than Combat Arms (N =2120)	Total (N = 4,506)
Combat Leadership (Cognitive)	.06	.05	.06
Technical-Managerial Leadership (Cognitive)	.02	-.01	.01
Career Potential (Cognitive)	.01	.04	.02*
Combat Leadership (Non-Cognitive)	.08*	.10**	.09**
Technical-Managerial Leadership (Non-Cognitive)	.10**	.09**	.10**
Career Potential (Non-Cognitive)	.02	.01	.02
Career Intent	.02	.06	.04
OBC Grades	.25**	.15**	.20**
Mid-Course Peer Ratings	.12**	.11**	.11**
Final Peer Ratings	.22**	.21**	.21**
Multiple Correlations	.30**	.24**	.26**

*Significant at the .05 level.
**Significant at the .01 level.

REFERENCES

Downey, R. G., Medland, F. F., & Yates, L. G. Evaluation of a peer rating system for predicting subsequent promotion of senior military officers. Journal of Applied Psychology, 1976, 61, 206-209.

Fleishman, E. A. Twenty years of consideration and structure. In Fleishman, E. A. and J. G. Hunt (Eds.) Current development in the study of leadership. Carbondale, Ill.: Southern Illinois University Press, 1974.

Gilbert, A. C. F. Prediction of Army officer performance. Paper presented at the Annual Conference of the Military Testing Association, Indianapolis, Ind., September 15-19, 1975. In Proceedings of the 17th Annual Conference of the Military Testing Association. Indianapolis, Ind.: U.S. Army Enlisted Evaluation Center, 1975.

Gilbert, A. C. F., & Downey, R. G. Validity of peer ratings obtained during Ranger training (Technical Report). Alexandria, Va: U.S. Army Research Institute for the Behavioral and Social Sciences, in preparation, 1977.

Gilbert, A. C. F. & Grafton, F. C. Characteristics of an officer performance evaluation measure. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, Ala., October 18-22, 1976. In Proceedings of the 18th Annual Conference of the Military Testing Association. Pensacola, Fla.: Naval Education and Training Program Development Center, 1976.

Gilbert, A. C. F., Hooper, J. N., & Hicks, J. M. Establishing the validity of a multidimensional Army officer performance measure. Paper presented at the Annual Meeting of the Southeastern Society for Multivariate Experimental Psychology, Hollywood, Fla., May 4-7, 1977.

Gordon, L. V., & Medland, F. F. The cross-group stability of peer ratings of leadership potential. Personnel Psychology, 1965, 18, 173-177.

Haggerty, H. R. Status report on research for the U.S. Military Academy (Technical Report 1133). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, October 1963.

Helme, W. H. Research to predict cadet and officer performance (Research Study 69-10). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, May 1969.

Helme, W. H., Willemin, L. P., & Grafton, F. C. Dimensions of leadership in a simulated combat situation (Technical Research Report 1172). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, July 1971. (NTIS No. AD 730 315)

Parrish, J. A., & Drucker, A. J. Personnel research for Officer Candidate School (Technical Research Report 1107). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, November 1957.

Stogdill, R. M. Handbook of leadership: A survey of theory and research. New York: The Free Press, 1974.

Willemin, L. P. Criterion aspects of Army research on the prediction of officer performance (Research Study 65-6). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, December 1965.

USING AN ASSESSMENT CENTER TO PREDICT FIELD LEADERSHIP PERFORMANCE OF ARMY OFFICERS AND NCOs

Frederick N. Dyer
Richard E. Hilligoss
Army Research Institute Field Unit
Fort Benning, Georgia

INTRODUCTION

The assessment center concept involves the immersion of individuals into situations which simulate those he would face if he were selected for entry or promotion. It has been widely used in industry and business to select personnel for high level positions. In 1973-1974 the U.S. Army Infantry School (USAIS) Assessment Center (ACTR) assessed students from the Infantry Officer Advanced Course (IOAC), the Infantry Officer Basic Course (IOBC) and the Advanced NCO Educational System (ANCOES) to determine the feasibility of the assessment center concept as a leadership development and leadership prediction technique. It also assessed students from the Branch Immaterial Officer Candidate Course (BIOCC) to determine the feasibility of the assessment center concept as a selection device. The purpose of the present paper is to discuss the effectiveness of the ACTR for predicting field leadership performance.

ACTR DESCRIPTION

Table 1 presents a summary of assessee characteristics and group sizes. Assessees reported to Fort Benning one week before their scheduled USAIS course to participate in the assessment center. Day I was sign in, room assignment, meal arrangements, etc; Day II, Day III and 1/2 of Day IV were given to the assessment process, providing 2 1/2 days of assessment. Two pilot sessions were completed in June-July 1973, with the first regular session beginning 11 July 1973.

The assessor pool consisted of six Majors, seven Captains, two Lieutenants, three Master Sergeants, two Sergeants First Class, and one Staff Sergeant. The assessors were selected by DA using the following criteria: each man must be in one of the combat arms, each Captain and above must have had command experience; each Major, Captain, and Sergeant must have served in combat, and Officers must have had an advanced degree in one of the behavioral sciences.

Table 1

ASSESSEE GROUP CHARACTERISTICS AND SIZES

	ASSESSMENT GROUP			
Descriptor	IOBC	IOAC	BIOCC(OCS)	ANCOES
Number Assessed	90	88	143	87
Number with complete 6-month ratings	45	36	40	38
Pay Grade	O 1	O 3	E 3-6	E 6-7
Average Age	22.6	28.8	25.3	33.3
Average years of Active Duty	0.3	5.7	3.3	12.9

LEADERSHIP DIMENSIONS

The staff of the assessment center and Army Research Institute and HumRRO scientists identified lists of behavioral dimensions characteristic of effective leadership. The basis of this selection was the Ohio State Leadership Study, The Army War College report, Leadership for the 70's, and other prominent studies on leadership.

The ten dimensions of leader behavior judged to be appropriate for the assigned mission and which it was believed could be evaluated using the assessment center concept were:

Adaptability
Administrative Skills
Communication Skills
Decision Making
Forcefulness
Mental Ability
Motivation
Effectiveness in an Organizational Leadership Role
Social Skills
Supervisory Skills

ASSESSMENT CENTER EXERCISES

The staff then began the selection and construction of exercises and questionnaires. In evaluating possible exercises and exercise concepts, a basic factor of consideration was that the exercises would place the assessees in uniquely different situations while simultaneously providing multiple opportunities for the evaluation of each dimension. Exercises were ultimately selected based upon their situational diversity, military relevance and apparent potential for eliciting behaviors related to the designated dimensions. The following battery of exercises were selected:

Entry Interview: A background interview to elicit information related to motivation, experience and the assessee's self-knowledge of his strengths and weaknesses (90').

Appraisal Interview: An applied exercise in which each assessee interviewed two others to select one for a position within a battalion. This interview elicited behaviors related to communication skills, social interaction and organization of thought (100').

Leaderless Group Discussion: This exercise was a combined individual and group task in which 6 IOAC assessees were assigned a mission to distribute year end funds among the represented directorates while attempting to acquire a maximum amount for his own directorate. IOBC, BIOCC, and ANCOES assessees were assigned a mission to get a soldier from their unit selected as the Brigade Soldier of the Month and providing a rank order of merit list of the available candidates. This exercise elicited behaviors associated with forcefulness, persuasiveness, organizational ability and group interaction (120').

In-Basket Exercise (Three versions: IOAC - assessee was placed in the role of a battalion commander; IOBC/BIOCC - assessee was placed in the role of a company commander; ANCOES - assessee was placed in the role of a 1st Sergeant). An in-basket containing many items typical of the appropriate position was presented to the assessee who had 3 hours to address each item in the in-basket. This exercise elicited behaviors relating to problem solving, decision making, work organization and leadership. It was followed by an interview to discuss reasons for action taken and the relationship perceived to exist among some of the actions (180').

War Game (IOAC assessees only): This was an assigned-role rotating leader exercise conducted in two 2-hour sessions. Teams of 6 players engaged in cost effectiveness analysis in a military force planning environment. Total costs, R&D, intelligence acquisition, balanced offensive/defensive forces were all considered under limited budget and time constraints. This exercise elicited organizational and leadership behavior (240').

Radio Simulate (Three versions: IOAC assessees were placed in company commander role; IOBC/BIOCC assessees were placed in a platoon leader role during a civilian emergency situation to insure that lack of military experience did not preclude them from participation in the exercises; ANCOES assessees were placed in the role of acting platoon leaders). It was a 5-hour exercise using radios as the only means of communication. It elicited organizational and leadership behaviors (300').

Field Exercise (IOBC, BIOCC, ANCOES): This was a 5-hour rotating leader designated exercise involving a team of 6 assessees. There were 6 lanes with a different obstacle provided for each lane. A separate mission and choice of materials was also available. It elicited emergent leadership, planning and organizational behaviors (300').

Management Exercise ("Conglomerate"): This was a two hour exercise divided into two planning and two trading periods. The 18-man assessment group was organized into three 6-man groups who competed against each other. This exercise elicited behaviors relating to emergent leadership, aggressiveness and social interaction (120').

Writing Exercise: This was an exercise designed to measure accuracy of information provided, grammar, spelling, and completeness. The IOAC assessees responded to a Staff Action Paper and the other assessment groups to a discharge action (60').

PSYCHOMETRIC TESTS AND SELF-DESCRIPTION INSTRUMENTS

A survey of tests in general was made revealing many possibilities for adoption into the assessment program. The primary criterion for selecting specific tests was relevance of the variables to be tested to the dimensions of leadership: administrative skills, communication skills, supervisory skills, forcefulness, adaptability, decision making, and mental ability.

The secondary criteria used in selecting tests were: non-offensive test items, suitability in content and format for use with mature adults, adequacy of normative data and theoretical discussions, recency of publication or revision and efficiency in test administration.

Both cognitive and non-cognitive tests were selected specifically to (1) allow for the comparison of an individual score with normative data and (2) verify the results of other assessment measurements. Group tests were selected in order to minimize the number of assessors and the amount of time required for each assessment. The psychometric tests and self-descriptive instruments selected are listed below.

1. Leadership Opinion Questionnaire
2. Watson-Glaser Critical Thinking Appraisal
3. Nelson-Denny Reading Test
4. Henmon-Nelson Test of Mental Ability
5. Leadership Q-Sort Test
6. Social Insight Test (Chapin)
7. Work Environment Preference Schedule (Gordon)
8. Strong Vocational Interest Blank
9. Edwards Personal Preference Schedule
10. Person Description Blank

Questionnaires to obtain specific background information about the assessee, and to solicit the assessee's opinion of his assessment experience, were also developed. The purpose of these questionnaires was to assist in the overall research effort and to collect suggestions for improving Assessment Center techniques and administration.

FEEDBACK TO ASSESSEES

To provide an impetus for individual self-development of leadership attributes, IOAC and ANCOES students were provided post-assessment counseling feedback. The purpose of this session was to enhance professional development through identifying the counselee's leadership strengths and weaknesses as observed at the Assessment Center. The counselor discussed with the counselee each behavioral area assessed, described the observations noted by assessors, and guided him towards available and effective materials that would be beneficial to him in developing a self-improvement program. The counselor's approach was essentially non-directive, as any actions for self-improvement were left entirely to the counselee.

FIELD LEADERSHIP PERFORMANCE RATINGS

The leadership criterion used to validate the ACTR measures consisted of ratings of ten leadership dimensions by two superiors, two peers and two subordinates of the assessee. These were made six months following the completion of the assessee's USAIS course by personnel in his new unit. The same ratings were again obtained 18 months following completion of school although fewer questionnaires were returned at this later period. Where ratings were obtained at both periods, there was only a 10% overlap in raters from the first period to the second.

The ten leadership dimensions were Decision Making, Administrative Skills, Interpersonal Competence (Social Skills), Communication Skills, Supervisory Skills, Organizational Role Skills, Technical and Tactical Competence, Leader Motivation, Leader Adaptability, and Leader Forcefulness. For each dimension five statements describing particular behaviors were rated making a total of 50 items on the Leadership Performance Rating Form (LPRF).

Approximately one-half of the questionnaires were returned. Complete rating data was obtained on 159 of the original 408

assessees at six months and complete data was obtained on 108 assessees at six and 18 months.

The average rating for all 50 items per questionnaire and all six questionnaires was calculated for six months and for 18 months. The correlations between these two averages ranged from .54 for the IOBC assessees, through .68 for IOAC assessees, to .75 for the ANCOES assessees. Only 15 BIOCC assessees had complete rating data for 6 and 18 months and the negative correlation between six and 18 month averages for this group (-.35) may have been a spurious result. The six-month/18-month correlation may be thought of as a test/retest reliability. These correlations are surprisingly high since many factors could operate to change leadership over the 12-month period between ratings and because of the relatively short time for observation of leadership prior to the first ratings (six months). Correlations between rater types (superior, peer, subordinate) were also generally significant and positive for each rating period (for IOBC, IOAC, and ANCOES assessees).

Although these correlations indicate the overall average rating at a rating period was highly reliable, the questionnaire failed to discriminate among the ten dimensions that presumably were represented in the fifty items. A factor analysis indicated only one significant factor which accounted for 74% of this common variance. It is not clear whether the failure to discriminate among leadership dimensions reflected on ratee performance or whether the different leadership dimensions are as interdependent as these high correlations indicate.

Since much more data was available for the six-month rating period (with almost no 18-month data from the BIOCC assessees) and since a high correlation existed where such data were available, the average rating for all 300 questions (six raters x 50 questions) at the six-month rating period was used as the field leadership criterion to validate the ACTR measures.

RESULTS

The scores obtained from the ACTR fall into the following six classes:

1. Assessor ratings of assessee performance during individual and group formal exercises such as the In-Basket,

2. Peer rankings of assessees in those formal exercises where a group of assessees participated together such as the Assigned Leader Group Exercise,

3. Self-rankings by the assessee of his performance relative to other group members in these group exercises,

4. Leadership dimension ratings made by an assessor during the Entry Interview with the assessee,

5. Assessee performance on paper and pencil performance tests, and

6. Assessee self-descriptions on questionnaires and other instruments such as the Edwards Personal Preference Schedule.

The results will be discussed for each of the above classes of score and, following this, the classes of ACTR scores themselves will be discussed and compared on their effectiveness for prediction of the field leadership ratings criterion. Proportions of successful predictors will be compared among classes as will the amount of time required by assessors and assessees to obtain each successful measure. The end result will be an ordering of the different classes of ACTR measure on their utility for predicting the criterion.

1. ASSESSOR RATINGS OF ASSESSEE PERFORMANCE DURING FORMAL EXERCISES

<u>Leaderless Group Discussion</u>

Assessor ratings for this exercise provided good predictors of the field leadership criterion for the IOBC assessee group. In particular, a rating of "amount of negative social behavior shown" was correlated ($r=-.56$, $p<.01$) with the criterion indicating that those assessees who showed more negative social behavior were more likely to be rated high on field leadership. Similarly, "social concern" was related to the criterion with IOBC assessees who showed less social concern being more apt to be rated high on field leadership ($r=-.37$, $p<.01$). One other rated dimension that was significantly related to the criterion for this group was "speaking ability". IOBC assessees who were rated high on this dimension were more apt to be rated high on field leadership ($r=.28$, $p<.05$).

For BIOCC assessees, "social concern" was significantly related to the criterion ($r=.31$, $p<.05$) but, contrary to IOBC, high social concern was related to good ratings on the criterion. "Amount of negative social behavior" showed a similar reversed relation to the criterion (compared to IOBC) although the correlation was not significant ($r=.24$, $p=.06$).

For ANCOES assessees, the Leaderless Group Discussion produced a single significant relation with the criterion. The dimension "conveys information" was correlated negatively ($r=-.32$, $p<.05$), indicating that persons rated lower on this communication skill dimension were more apt to be rated high on the criterion. As will be shown throughout this section, poor performance for NCOs on the ACTR exercises was frequently related to higher ratings on the criterion.

Assessor ratings on the Leaderless Group Discussion failed to predict the criterion for the IOAC assessee group.

<u>Conglomerate Exercise</u>

Only two of the assessor ratings for this exercise showed significant relationships with the criterion. For the IOBC assessees, ratings of "energy and vigor" were negatively correlated ($r=-.26$, $p<.05$) indicating that low energy and vigor were more apt to be related to high field leadership ratings. For the BIOCC assessees, the "receptivity" rating showed a positive correlation with the criterion ($r=.36$, $p<.01$). Assessees who were rated higher on "listening to and considering ideas of others" were more apt to receive high field leadership ratings.

Assessor ratings on the Conglomerate Exercise failed to predict the criterion for the ANCOES and IOAC groups.

<u>Radio Simulate</u>

Assessor ratings on the Radio Simulate exercise were almost completely unrelated to the field leadership criterion. Only for the ANCOES assessees were significant relations found for ratings of "communication skills" ($r=-.27$, $p<.05$), "adaptability" ($r=-.28$, $p<.05$) and "organizational identification" ($r=-.41$, $p<.01$). In each of these cases, poor NCO performance on the exercise was related to high criterion performance.

In-Basket

Assessor ratings on this exercise showed significant relations to the criterion for all groups but the IOBC assessees. For IOAC captains, the field leadership criterion was positively related to good assessor ratings on "decision making" ($r=.33$, $p<.05$) and a "use of available information" ($r=.36$, $p<.05$). For BIOCC assessees high criterion ratings were related to good performance on "planning and organization" ($r=.27$, $p<.05$) and "task orientation" ($r=.35$, $p<.05$).

All significant relationships between In-Basket assessor ratings and ANCOES field leadership ratings were negative. Good criterion ratings were related to poor "directing ability" ($r=-.27$, $p<.05$), poor "task orientation" ($r=-.37$, $p=.01$) and poor "sensitivity" ($r=-.27$, $p=.05$).

Appraisal Interview

No assessor rating was significantly related to the criterion for the IOBC, IOAC and BIOCC assessee groups for this exercise. For the ANCOES assessee group only one dimension, "ability to organize" was related ($r=-.33$, $p<.05$). The negative correlation indicates that poor "ability to organize" on the exercise was related to good field leadership ratings.

Writing Exercise

Assessor ratings on "accuracy of written information" were significantly related to the criterion for both the IOBC and IOAC groups ($r=-.27$ and $r=-.29$, respectively, $p<.05$ for both). The negative relationship indicates that poorer writing accuracy was related to better field leadership ratings. The other significant relationship for this exercise was "spelling" which for the ANCOES assessees was related positively to the criterion ($r=.28$, $p<.05$).

Assessor ratings on the Writing Exercise failed to predict the criterion for the BIOCC group.

Assigned Leader Group Exercise (ALGE)

All assessee groups except the IOAC captains completed this exercise. This exercise was successful in predicting the criterion for the ANCOES group. High assessor ratings on two dimensions were associated with high field leadership ratings. These were "emergent leadership" ($r=.29$, $p<.05$) and "group

facilitation" ($r = .29$, $p < .05$). Interestingly, these were the two dimensions on the exercises that were classed as "follower behaviors". The other significant relationship indicated that low assessor ratings on "flexibility" were associated with high scores on the criterion ($r = -.30$, $p < .05$).

The ALGE assessor ratings provided no significant correlations with the criterion for the remaining IOBC and BIOCC assessee groups.

Leader Game

Only the IOAC captains participated in this exercise (it took the place of the ALGE for this group) and one of the assessor-rated dimensions "flexibility" was successful in predicting the criterion. Assessees with high flexibility ratings were apt to be rated high on field leadership ($r = .36$, $p < .02$). Among the nonsignificant assessor ratings, dimensions of "leadership", "planning", and "organization" which would be expected to have even stronger relations to a leadership criterion did not even approach significance.

2. PEER RANKINGS ON GROUP EXERCISES

Leaderless Group Discussion

The six group members who participated in this exercise ranked all six members on a number of different dimensions at the end of the exercise. No significant predictors of the criterion were found for any of the dimensions on which peer rankings were made.

Conglomerate Exercise

Similar rankings were obtained from group members in this exercise with similar results, i.e., no significant relationships with the criterion for any assessee group.

Assigned Leader Group Exercise

More predictive validity was found for peer rankings in this exercise. In fact three of the four dimensions provided significant criterion predictors for the ANCOES assessee group. These were "ability to lead" ($r = .29$, $p < .05$), "quality of leader support" ($r = .27$, $p < .05$), and "generating group morale" ($r = .33$,

$p<.05$). These positive correlations indicate that high-ranked individuals on the exercise tended to receive the high field leadership ratings. The only other significant correlation for this exercise appeared for the BIOCC assessee group for a ranking of "how much you would like to associate with them socially" ($r=.30$, $p<.05$). Persons preferred for socialization were more apt to be rated high on the criterion.

Leader Game

This exercise did not produce any significant peer ranking correlations with the criterion for the IOAC assessees who participated in it.

3. SELF RANKINGS ON GROUP EXERCISES

Leaderless Group Discussion

The assessee included himself in the group rankings for this exercise and his self-ranking was tested also as a predictor of the criterion. Only one of these scores was found to predict the criterion. This was the self-ranking on "idea quality" ($r=.32$, $p<.05$) for the ANCOES assessees. Persons who ranked themselves higher on this dimension were more apt to receive high field leadership ratings.

Conglomerate

Three self rankings were significantly associated with the criterion on this exercise for the ANCOES assessees. These were "popularity" ($r=.29$, $p<.05$), "energetic support of team effort" ($r=.34$, $p<.05$), and "causing conflict within the group" ($r=.29$, $p<.05$). High "popularity", high "energetic support of team effort" and low "amount of conflict" were related to high ratings of field leadership. For the IOAC group, self-rankings of "idea quality" were related positively to the criterion ($r=.31$, $p<.05$). IOBC and BIOCC assessees did not produce significant self-ranking predictors for this exercise.

Assigned Leader Group Exercise

The ANCOES assessee group produced the only significant self-ranking predictors for this exercise. These were for dimensions of "ability to lead" ($r=.32$, $p<.05$) and "generating group morale" ($r=.30$, $p<.05$). The positive correlations indicate high self-rankings were related to good ratings on the field leadership

criterion. IOBC assessees did not produce significant self-ranking predictors and the IOAC did not participate in this exercise.

Leader Game

As with peer rankings, self rankings produced no significant correlations with the criterion for the IOAC assessees who were the only participants of this exercise.

4. ENTRY INTERVIEW PERFORMANCE EVALUATION

Six of the 14 scores of the Entry Inverview significantly predicted the field leadership ratings of the BIOCC assessee group. These were "overall impression" ($r=.42$, $p<.01$), "interest in self-development" ($r=.28$, $p<.05$), "effectiveness in conveying information" ($r=.35$, $p<.05$), "derives satisfaction from work accomplishments" ($r=.31$, $p<.05$), "fluent and articulate" ($r=.29$, $p<.05$), and "how well he expresses his opinions" ($r=.29$, $p<.05$). These positive correlations indicate that good Entry Interview ratings were related to good field leadership criterion ratings.

The ANCOES assessees who were rated high on "animation and enthusiasm" were much more apt to receive high criterion ratings than their lower-rated colleagues ($r=.45$, $p<.01$). For this group "interest in self-development" was inversely related to the field leadership ratings ($r=-.29$, $p<.05$). The only other significant predictor from the Entry Interview was for the IOBC group. As for the ANCOES group, "interest in self-development" was negatively correlated with field ratings of leadership ($r=-.27$, $p<.05$).

5. PENCIL AND PAPER PERFORMANCE TESTS

The four tests that fall into this category are the Henmon-Nelson Tests of Mental Ability, the Watson-Glaser Critical Thinking Appraisal, the Nelson-Denny Reading Test, and the Social Insight Test. Only for the ANCOES assessee group did these measures successfully predict the field leadership ratings criterion. However, it is questionable to use the term "successfully" since poor performance on the Henmon-Nelson Quantitative ($r=-.30$, $p<.05$); Henmon-Nelson Verbal ($r=-.41$, $p<.01$), Henmon-Nelson Total Score ($r=-.40$, $p<.01$); Nelson Denny Vocabulary ($r=-.36$, $p<.05$); Nelson-Denny Comprehension ($r=-.32$, $p<.05$) and Nelson Denny Total ($r=-.37$, $p<.05$) were related to good ratings on the field leadership criterion. The Watson Glaser Critical Thinking Appraisal and The Social Insight Test showed no

significant correlations with the criterion for any of the assessee groups.

6. SELF-DESCRIPTION INSTRUMENTS

Edwards Personal Preference Schedule (EPPS)

One of the highest correlations obtained with the criterion was from this instrument. IOAC assessees with a high "Need for Order" tended to be rated higher on the field ratings of leadership ($r=.52$, $p<.001$). In addition, the IOAC assessees showed an inverse relationship between "Need for Succorance" (to have others provide help when in trouble, to seek encouragement from others, etc.) and the criterion ($r=-.35$, $p<.05$).

The ANCOES assessee group also showed a number of significant correlations between EPPS measures and the criterion. "Need for Exhibition" was inversely related to the criterion ($r=-.31$, $p<.05$), and "Need for Abasement" was related positively ($r=.28$, $p<.05$). Finally scores on the "Consistency" variable were related to the criterion: positively for the IOBC assessees ($r=.30$, $p<.05$) and negatively for the ANCOES group ($r=-.34$, $p<.05$). No EPPS measures were significantly related to the criterion performance of the BIOCC assessees.

Work Environment Preference Schedule (WEPS)

High scores on this measure "typify individuals who accept authority, who prefer to have specific rules and guidelines to follow, who prefer impersonalized work relationships, and who seek the security of organizational and in-group identification." Two of the assessee groups showed significant correlations of their scores on this measure with their criterion field leadership ratings. IOBC assessees who were lower on the WEPS were more likely to receive high criterion ratings ($r=-.25$, $p<.05$) and IOAC assessees who were higher on the WEPS were more likely to receive high criterion ratings ($r=.32$, $p<.05$). The BIOCC and ANCOES groups did not have significant correlations with the criterion on this measure.

Leader Opinion Questionnaire (LOQ)

ANCOES assessees scoring high on "Consideration" on the LOQ were more apt to be rated high on the criterion ($r=.36$, $p<.05$). IOBC assessees who were high on "Structure" were more apt to be

rated high on the criterion ($r=.25$, $p<.05$). No other LOQ scores were significant for these or for the other assessee groups.

Leadership Q Sort (LQS)

IOBC assessees showed a fairly strong relationship of "Decision Making" to the criterion with the persons scoring low on this dimension being more apt to receive high leadership ratings ($r=-.39$, $p<.01$). "Teaching and Communication" scores, on the other hand were positively related to the criterion for the IOBC group ($r=.27$, $p<.05$). High scores on "Mental Health" were related to high criterion ratings for the ANCOES assessees ($r=.33$, $p<.05$) while low scores on "Personal Integrity" were related to high criterion ratings for this group ($r=-.30$, $p<.05$).

IOAC assessees showed an inverse relation between "Consideration" scores and the criterion ($r=-.36$, $p<.05$). BIOCC assessees showed no significant relationship of LQS measures to the criterion.

Person Description Blank

Fifty pairs of adjectives were presented to each assessee (e.g. WARY: 1 2 3 4 5 6 7: GULLIBLE) with instructions to rate himself by circling the number that best described his position between these polar adjectives. Twenty-six of these fifty pairs produced significant correlations with the criterion for at least one of the assessee groups. The pairs of adjectives and their correlations with the criterion for each assessee group are presented in Table 2. Positive correlations indicate that persons who rated themselves higher than average on the rightmost adjective were more apt to be rated high on field leadership. Negative correlations indicate that persons who rated themselves higher than average on the leftmost adjective were more apt to be rated high on field leadership. A negative correlation does not necessarily mean that people were closer to the "1" end of the scale than to the "7" end of the scale. It only indicates that persons who were on the "1" side of the overall average for that item were more apt to be rated high on the criterion.

COMPARISON OF DIFFERENT CLASSES OF ACTR SCORES

Table 3 presents summary data for all assessee groups for the six classes of ACTR scores. It can be seen that the number of scores per assessee (Column 1) varied from 9 for the Pencil and Paper Performance Tests to 75 for the Self-Description

Table 2

PERSON DESCRIPTION BLANK (PDB) "YOURSELF" SCORE

CORRELATIONS WITH CRITERION

PDB Descriptor	Assessment Group IOAC	IOBC	BIOCC	ANCOES
Noncompetitive (1) competitive (7)	.26	.45(.001)**	.25	-.17
Clumsy (1) Graceful (7)	.19	.20	.31(.026)*	.12
Understandable (1) Mysterious (7)	-.32(.029)*	.03	.13	-.09
Insensitive (1) Sensitive (7)	-.02	.40(.004)**	.12	-.05
Yielding (1) Firm (7)	.14	.29(.026)*	-.10	.41(.005)**
Tough (1) Tender (7)	-.39(.009)**	-.17	.17	-.31(.030)*
Military (1) Unmilitary (7)	-.20	.16	.18	-.28(.042)*
Wary (1) Gullible (7)	-.49(.001)**	-.45(.001)**	-.11	-.19
Indecisive (1) Decisive (7)	.07	.27(.035)*	-.12	.16
Careful (1) Reckless (7)	-.16	-.10	-.01	-.38(.009)**
Cowardly (1) Brave (7)	-.04	.04	-.13	.33(.023)*
Insincere (1) Sincere (7)	-.11	.15	-.39(.006)**	.16

* .05, ** .01

Table 2 (cont'd)

PERSON DESCRIPTION BLANK (PDB) "YOURSELF" SCORE

CORRELATIONS WITH CRITERION

	Assessment Group			
PDB Descriptor	IOAC	IOBC	BIOCC	ANCOES
Leading (1) Following (7)	-.24	-.31(.020)*	-.04	-.15
Passive (1) Active (7)	.34(.022)*	.15	-.11	.22
Soothing (1) Irritating (7)	.03	-.24	-.18	-.39(.008)**
Ambitious (1) Complacent (7)	-.36(.016)*	-.19	-.10	-.12
Boring (1) Interesting (7)	-.22	.18	.03	.27(.048)*
Quiet (1) Talkative (7)	-.31(.031)*	.06	-.11	.07
Secretive (1) Open (7)	-.34(.021)*	.02	-.28(.041)*	.18
Mission-oriented (1) People-Oriented (7)	-.32(.028)*	.19	.21	-.19
Colorful (1) Colorless (7)	.12	-.18	-.13	-.32(.025)*
Hard-working (1) Easy-Going (7)	-.44(.003)**	-.12	-.10	-.25
Dominating (1) Submissive (7)	-.29(.043)*	.03	.12	-.13
Stable (1) Changeable (7)	-.33(.023)*	.03	.07	.06
Unathletic (1) Athletic (7)	.20	.20	.31(.025)*	.44(.003)**
Disorganized (1) Organized (7)	.20	.37(.006)**	.02	.18

* .05, ** .01

Table 3: RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORES - ALL ASSESSEE GROUPS COMBINED

Class of ACTR Score	No. of Scores	Average Number Successful Predictors	% Successful Predictors	Assessor time per Assessee Score (min)	Assessor time per successful predictor (min)	Assessee time per score (min)	Assessee time per successful predictor (min)
Assessor Ratings Formal Exercises	69	6.5	9.4	10.9	115.9		
Peer Rankings Formal Exercises	15	1	6.7	0	0	12.1	129.5
Self Rankings Formal Exercises	15	1.75	11.7	0	0		
Entry Interview	14	2	14.3	4.6	32.5	4.6	32.5
Pencil & Paper Performance Tests	9	1.5	16.7	4.3	25.8	17.8	106.7
Self-Description Instruments	75	11.5	15.3	.30	1.99	1.83	12.0

Instruments. The assessor time per score (Column 4) showed a very wide variation from 10.9 minutes per score for Assessor Ratings on Formal Exercises to less than one minute per score for the Self-Description Instruments. The latter small time per score reflects the assessor time savings that resulted from presenting the Self-Description Instruments in a group (six assessees) setting. The zero "assessor times per score" that appear for Peer Rankings and Self Rankings reflect the fact that these scores were provided by the assessees and did not require any additional time of assessors beyond that required for the assessor ratings on these exercises. The "assessee time per score" (Column 6) is prorated over Assessor Ratings, Peer Rankings and Self Rankings. Thus only a single figure is shown for this column for these three categories. It can be seen that assessee time per score is also long for the Formal ACTR Exercises. Assessee time per score is longest for the Pencil and Paper Performance Tests and shortest for the Self-Description Instruments.

A successful predictor is defined in this report as one which has a correlation with the criterion that is significant at the .05 level. In Column 2 of Table 3 the average number of successful predictors per assessee is given and Column 3 shows the percentage that this is of the total number of scores for the assessee. Five percent would be expected by chance due to the .05 significance level. This figure ranges from a high of 16.7% for the Paper and Pencil Performance Tests to near chance levels (6.7%) for the Peer Rankings. The high figure for the Pencil and Paper Performance Tests is somewhat misleading since all of the significant predictors were for the ANCOES group and all indicated poor pencil and paper test performance to be related to good field ratings (see below). Perhaps the most interesting data is in Column 5 where the assessor time per successful predictor for each class of ACTR score is shown. This ranges from 2 minutes per successful predictor for the Self-Description Instruments to nearly two hours per such predictor for the Assessor Ratings of Formal Exercises.

The assessor ratings of formal exercises represent the most typical ACTR data and their collection is the raison d'etre of an assessment center. The poor predictions from these rating scores compared to interviews, and to questionnaires is thus especially disappointing for ACTR proponents. The poor performance is not a result of low rating reliability. Checks of rater reliability on the exercises where more than one assessor rated the same assessee indicated that reliability of the ratings was surprisingly good. Spearmon-Brown calculations indicate the three-rater sums for LGD,

ALGE, CONG and LGAM to have reliabilities in the 70s and 80s.

The high reliability of the criterion field leadership ratings was described earlier. Since both criterion and assessor ratings are reliable, the failure of the assessor ratings to provide more than a few significant correlations with the criterion must reflect some failure of the ACTR exercises to elicit and/or measure the same behaviors that peers, superiors, and subordinates in field units classify as "leadership".

Tables 4, 5, 6 and 7 provide the data of Table 3 with a separate breakdown by the different assessee groups. It can be seen that the ANCOES scores (Table 7) provide much better prediction of the criterion than the ACTR scores of any of the other assessee groups. However, a sizable portion of the significant ANCOES criterion predictors represent a troublesome inverse relation between ACTR performance and the criterion. One normally would not intentionally set up an ACTR with the intent of selecting for promotion or employment only those persons who do badly on the ACTR tasks. These inverse relationships between predictor and criterion reflect a failure of the ACTR exercises, the unsuitability of the criterion, or both - at least for the ANCOES group.

Another result that is apparent from Tables 4, 5, 6 and 7 is that different assessee groups often have different patterns of success for the different classes of ACTR scores. For example, the Entry Interview does an excellent job for the BIOCC group (36% successful predictors) but it does little predicting for any other group. For IOAC assessees, the Self-Description Instruments do a good job of predicting the criterion but the other classes of score have little predictive validity.

Table 8 represents a breakdown of the data in Table 3 by separate exercise. The most effective single measure by almost all criteria is the Person Description Blank. This instrument required less than ten minutes to administer but provides much more effective criterion prediction than exercises such as the Radio Simulate which required five hours of assessee time, and even more assessor time. However, it can be argued that self-descriptions would be much less effective in a setting where deliberately falsified self-descriptions might occur. False self-descriptions would have been at a minimum in the USAIS ACTR since the assessees were assured that the data would not affect their careers.

Table 4

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE: IOBC ASSESSEES

Descriptor	No. of Scores	Number of Successful Predictors	% Successful Predictors	Assessor time per Assessee Score (min)	Assessor time per successful predictor (min)	Assessee time per Score (min)	Assessee time per successful predictor (min)
Assessor Ratings	69	5	7.2	10.9	151.4		
Peer Rankings	15	0	0	0	-	12.1	239.6
Self Rankings	15	0	0	0	-		
Entry Interview	14	1	7.1	4.6	64.8	4.6	64.8
Pencil & Paper Performance Tests	9	0	0	4.3	∞	17.8	∞
Self-Description Instruments	75	12	16.0	.30	1.88	1.83	11.4

Table 5

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE: IOAC ASSESSEES

Descriptor	No. of Scores	Number of Successful Predictors	% Successful Predictors	Assessor time per Assessee Score (min)	Assessor time per successful predictor (min)	Assessee time per Score (min)	Assessee time per successful Predictor (min)
Assessor Ratings	69	4	5.8	10.9	187.9		
Peer Rankings	15	0	0	0	-	12.1	239.6
Self Rankings	15	1	6.7	0	0		
Entry Interview	14	0	0	4.6	∞	4.6	∞
Pencil & Paper Performance Tests	9	0	0	4.3	∞	17.8	∞
Self-Description Instruments	75	15	20.0	.30	1.5	1.83	9.15

Table 6

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE: BIOCC ASSESSEES

Descriptor	No. of Scores	Number of Successful Predictors	% Successful Predictors	Assessor time per Assessee Score (min)	Assessor time per successful predictor (min)	Assessee time per Score (min)	Assessee time per successful Predictor (min)
Assessor Ratings	69	5	7.2	10.9	151.4	12.1	149.7
Peer Rankings	15	1	6.7	0	0		
Self Rankings	15	0	0	0	-		
Entry Interview	14	5	3[illegible].7	4.6	12.9	4.6	12.9
Pencil & Paper Performance Tests	9	0	0	4.3	∞	17.8	∞
Self-Description Instruments	75	4	5.3	.30	5.7	1.83	34.5

Table 7

RESULTS FOR SIX DIFFERENT CLASSES OF ACTR SCORE:AMCOES ASSESSEES

Descriptor	No. of Scores	Number of Successful Predictors	% Successful Predictors	Assessor time per Assessee Score (min)	Assessor time per successful predictor (min)	Assessee time per Score (min)	Assessee time per successful Predictor (min)
Assessor Ratings	69	12	17.4	10.[illegible]	62.6	12.1	57.6
Peer Rankings	15	3	20.0	0	0		
Self Rankings	15	6	40.0	0	0		
Entry Interview	14	2	14.3	4.6	32.2	4.6	32.2
Pencil & Paper Performance Tests	9	6	66.7	4.3	6.45	17.8	26.7
Self-Description Instruments	75	15	20.0	.30	1.5	1.83	9.15

Table 8

RESULTS FOR SEPARATE ACTR EXERCISES FOR ALL ASSESSEE GROUPS

Descriptor	No. Scores	Avg. No. Success Predictors	% Successful Predictors	Assessor time per Assessee Score (min)	Assessor time per Successful Predictor (min)	Assessee time per Score (min)	Assessee Time per Successful Predictor (min)
Assessor Ratings							
Leaderless Grp. Discussion	9	1.5	16.7	6.67	40.0		
Conglomerate	8	.5	6.2	7.5	121.0		
Radio Simulate	16	.75	4.7	15.0	319.1	15.0	100
In-Basket	15	1.75	11.7	7.3	62.4	15.3	130.8
Appraisal Interview	8	.25	3.1	12.8	413.0	16.3	127.3
Writing Exercises	4	.75	18.7	8.3	44.4	15.0	80.2
Assigned Leader Group Exercise & War Game (IOAC)	9	1.0	11.1	16.7	150.2		
Peer Ranking							
LGD	6	0	0	0	-	5.7[*a]	68.7[*a]
Cong	5	0	0	0	-	6.7[*a]	80.7[*a]
ALCE & War Game	4	.75	18.8	0	0	17.6[*a]	133.3[*a]
Self-Ranking							
LGD	6	.25	4.2	0	0		
Cong	5	1.0	20.0	0	0		
ALCE & War Game	4	.5	12.5	0	0		
Entry Interview	14	2	16.7	4.6	32.5	4.6	32.5
Performance Tests							
Henmon Nelson	3	.75	25	2.2	8.8	13.3	53.2

Table 8 (cont'd)

RESULTS FOR SEPARATE ACTR EXERCISES FOR ALL ASSESSEE GROUPS

Descriptor	No. Sorts	Avg No. Success Predictors	% Successful Predictors	Assessor Time Per Assessee Score (min)	Assessor Time Per Successful Predictor (min)	Assessee Time Per Score (min)	Assessee Time Per Successful Predictor (min)
Nelson-Denny	4	.75	18.8	1.7	9.0	10.0	53.2
Watson-Glaser	1	0	0	8.3	∞	30	∞
Social Insight	1	0	0	5.0	∞	30	∞
Self-Description Instruments							
Edwards Personal Preference Schedule	15	1.5	10	.56	5.6	3.33	33.3
Work Environment Preference Schedule	1	.5	50	1.67	3.33	10.0	20.0
Leadership Opinion Questionnaire	2	.5	25	1.67	6.67	10.0	40.0
Leadership Q Sort	7	1.25	17.6	1.19	6.74	7.14	40.6
Person Description Blank	50	7.75	15.5	.02	.13	0.14	.9

a* Peer and self-rankings included with assessor ratings for these calculations.

DISCUSSION

Two perspectives exist for discussion of these results. One is in terms of the specific characteristics as measured in the ACTR which predict field leadership ratings of the different assessee groups. The other perspective for viewing these results is in terms of the general question of what parts of the ACTR were effective in assessment of leadership.

CHARACTERISTICS OF SPECIFIC ASSESSEE GROUPS

The young lieutenant who had recently been commissioned and who, following his Infantry course was rated high on leadership by peers, subordinates and superiors, judged himself to be more wary, competitive, soothing and leading. His decision-making skills were rated lower by himself and by trained assessors. Ironically, he was judged to be somewhat lower on self-development than the lieutenant who was rated more poorly on field leadership.

The captain who was about to enter the Advanced Infantry Course and who later received high ratings on the field leadership criterion was apt to be high on his need for order and more apt to prefer a structured work environment. He performed well on in-basket exercises and viewed himself as more hard-working, wary, interesting, tough, ambitious, active, secretive, and stable.

The enlisted man about to enter Officer Candidate School and who, following his OCS training and Branch leadership course, was rated high on field leadership, was more apt than his low-rated colleague to make a good impression and to be fluent, creative and task-oriented. He viewed himself as more creative and persistent, yet somewhat less dominating and less sincere than his colleague who fared less well on field leadership ratings.

The NCO about to enter the advanced NCO course who later receives high ratings of field leadership was more enthusiastic but poorer in reading, quantitative and verbal skills than his colleague who received lower field leadership ratings. He was more considerate, but less able to perform on in-basket exercises and in simulated emergencies. He viewed himself as more athletic, firm, careful, soothing and brave than did his low-rated colleague.

PREDICTIVE VALIDITY OF DIFFERENT CLASSES OF ACTR SCORES

Self-Description Instruments provided the largest proportion of criterion predictors and also provided these scores with the least assessor and assessee time. On the other hand, the most assessor-intensive formal ACTR exercises actually do the poorest job of predicting the field leadership criterion. Intermediate between these extremes is the Entry Interview which provided a fair number of predictors with only a moderate amount of assessor and assessee time.

These results must be somewhat distressing to proponents of the assessment center concept. Such formal exercises as the In-Basket, Assigned Leader Group Exercise and Leaderless Group Discussion are the backbone of such centers. For such exercises to predict poorly in the current setting, despite good to excellent reliability of predictor and criterion measures, indicates a mismatch between the ACTR exercise measures and the criterion scores. A possible explanation of this mismatch is that the ACTR was more effective in eliciting leadership skills than the subsequent duties of these leaders. The USAIS ACTR exercises probably did provide tough challenges to leadership and actual assessee leadership skills were probably demonstrated for assessors to rate. However, the criterion ratings were made during peacetime when few if any emergencies would arise which required excellent leadership for their successful resolution. As a result, the criterion ratings may have been made on some other factor than leadership such as sociability. Another possible basis for field ratings may have been the leadership self-conceptions that the assessees held and somehow communicated to the superiors, peers, and subordinates who provided the criterion ratings. With few if any opportunities for assessees to demonstrate genuine leadership, this "talk about leadership" may have been the basis for leadership ratings. Not only would this account for the general failure of assessor-intensive exercises to predict the criterion, it would explain the relative success of instruments such as the Person Description Blank which were specifically designed to obtain leadership-related self conceptions.

Future validation studies planned for the USAIS ACTR assessees will utilize promotion data as a leadership criterion. Hopefully, promotions of these leaders would be related to their actual leadership skills and not to sociability or to their incorrect self-perceptions of their leadership skills.

ASSIGNMENT PROCEDURES IN THE AIR FORCE PROCUREMENT MANAGEMENT INFORMATION SYSTEM

Joe H. Ward, Jr.
Donald L. Haney
William H. Hendrix
Manuel Pina

Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Brooks AFB, Texas

I. INTRODUCTION

In July 1973, personnel from the Air Force Recruiting Service and Air Force Human Resources Laboratory (AFHRL) discussed strategies for examining the feasibility of a computer-based enlistment reservation system to enhance the existing Air Force Procurement Management Information System (PROMIS). A small computer-based job reservation system was developed using System 2000 data management system to demonstrate to recruiting service personnel the feasibility of on-line job reservations (Ward and Haltman, 1975). This demonstration, in September 1973, resulted in the development by Air Force Military Personnel Center, Recruiting Service and AFHRL of an operational job-reservation system (Pina and Stifle, 1977). The system became operational 1 November 1976, with Air Force representatives at the sixty-six Armed Forces Examining and Entrance Stations (AFEES) inquiring through remote terminals to a Burroughs 6700 computer located at Randolph AFB, Texas.

This paper discusses: (1) designing personnel systems for acceptance and improvement, (2) a general framework for viewing personnel assignment systems, (3) the procedure for offering jobs in the PROMIS system

II. DESIGNING PERSONNEL SYSTEMS FOR ACCEPTANCE, EVOLUTIONARY IMPROVEMENT, AND TECHNOLOGY TRANSFER

A Personnel System may be viewed as a vehicle to aid in improving the effectiveness of an organization. To be useful, a Personnel System should be <u>designed</u> for:

DESIGNING PERSONNEL SYSTEMS FOR ACCEPTANCE AND IMPROVEMENT

- ACCEPTANCE BY MANAGERS AND MEMBERS OF THE ORGANIZATION
- EVOLUNTIONARY (INCREMENTAL) ADJUSTMENTS LEADING TO CONTINUED IMPROVEMENT
- EASE OF INCORPORATING NEW HUMAN RESOURCES RESEARCH FINDINGS INTO THE OPERATIONAL PERSONNEL SYSTEM

Acceptance

If a personnel system is to have an opportunity to help an organization, it must continue to exist. In order to exist, it must be acceptable to managers and members of the organization. Designers of a personnel system must plan for initial and continued acceptance by members of the organization.

Evoluntionary Improvement

Designers of a personnel system must allow for future changes--both expected and unexpected. The system should expect those future policy changes designed to improve personnel effectiveness. However, it is impossible to foresee the problems that can arise after operational implementation. The design features of the system that allow for change also help insure continued acceptance. The capability to change must be approached with caution, since too frequent or too much change might lead to non-acceptance and destruction of the personnel system.

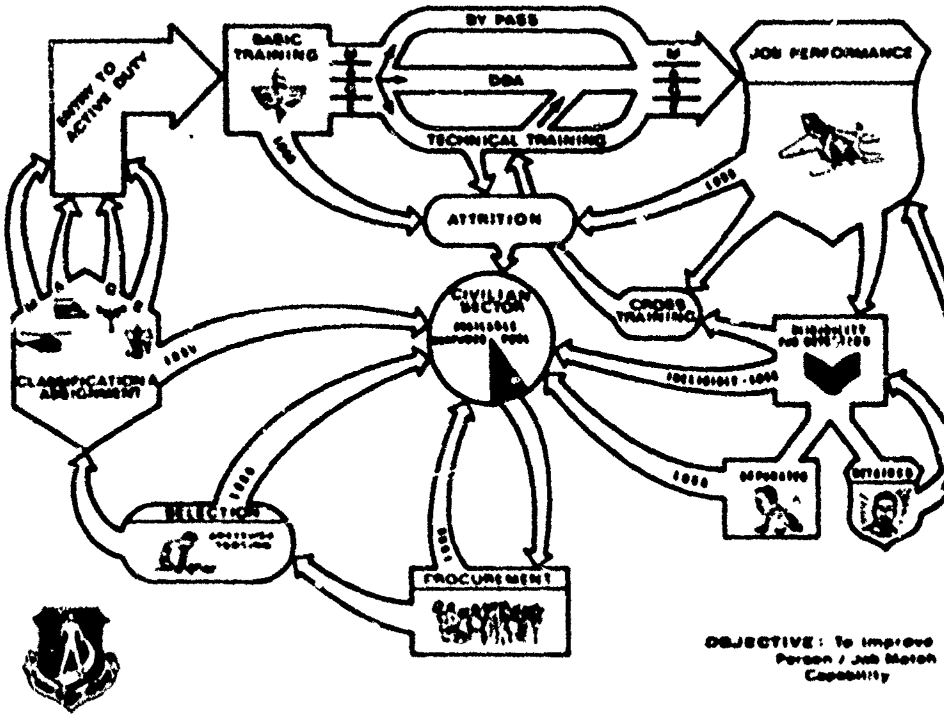
Incorporating New Research

In addition to allowance for expected management changes and unexpected problems, it is highly desirable to design a personnel system for acceptance of new human resources research findings. Some new technologies may require major modifications to the system. However, many future improvements can be incorporated easily into the operational system if it contains a technology transfer capability.

III. A VIEW OF PERSON-JOB ASSIGNMENTS

This section presents a view of person-job assignments that allows for user acceptance, evoluntionary improvement, and transfer of new research findings. The concepts to be described emphasize information about jobs and people, pay-off or utility of particular person-job assignments, and the contribution of each particular assignment to overall system effectiveness. Before examining the details, it is helpful to look at the Military Career Life Cycle.

MILITARY CAREER LIFE CYCLE



This picture represents some of the personnel decision activities that take place during a military career. The objective is for persons to move through various job or training activities so that overall system effectiveness is maximized. The following ideas reflect some essential features of a personnel system designed for acceptance and improvement.

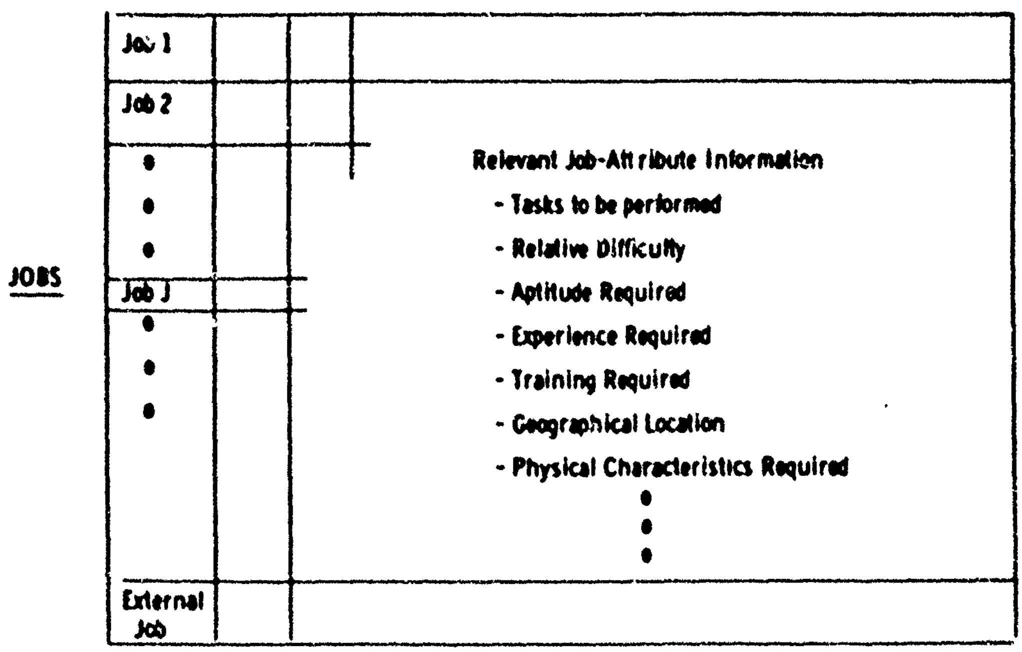
Activities to be Accomplished (Job and Training Requirements)

A necessary first step is the determination of the kinds of activities (jobs or training) that must be performed in the Air Force. This will be done from information about training requirements, job requirements, occupational surveys, and other sources. The attributes associated with jobs (or training positions) will be called job properties. Figure 1, the JOB PROPERTIES ARRAY, represents the relevant job-attribute information that is used in the personnel assignment system. The word JOB refers to any descriptive state of being that is occupied by or is potentially occupied by a person. The general term "jobs" can include

all Air Force jobs, plus activities that might be termed "training jobs." Another important "job" concept is the last one shown in Figure 1--called an External Job. This category provides for a job outside the particular sub-system of interest. The inclusion of an External Job allows for rejecting personnel by assignment to a "job" outside the system. In the Advanced Personnel Data System, Procurement Management Information System (APDS-PROMIS) each applicant occupies an External Job prior to assignment to an Air Force job.

Figure 1

JOB PROPERTIES ARRAY



Personnel Required to Accomplish the Activities

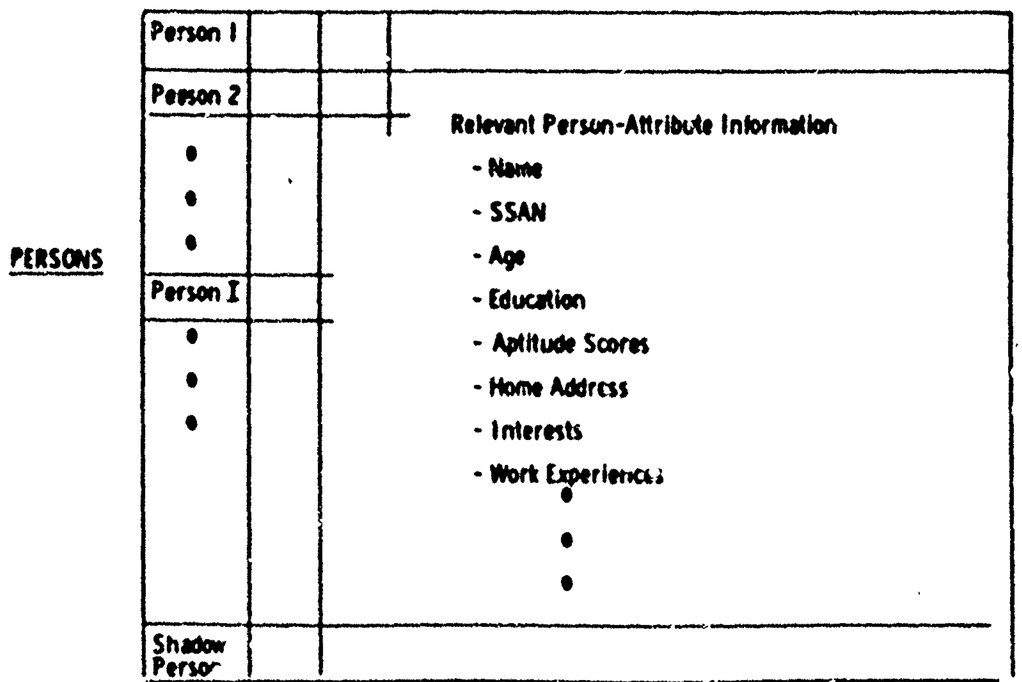
After the jobs have been determined it is necessary to identify the personnel that are available or potentially available to accomplish the activities required to operate the Air Force. The attributes associated with persons will be called person characteristics. Figure 2, PERSON CHARACTERISTICS ARRAY, represents the relevant person-attribute information that is used in the personnel assignment system. The word PERSON

refers to any individual that is a member of the Air Force or is a potential member of the Air Force.

The last person indicated in Figure 2 is called a Shadow Person. This Shadow Person provides for an imaginary person to be considered for assignment. The inclusion of this Shadow Person allows for consideration of Air Force jobs that are unfilled. The consequences of unfilled jobs (represented by assigning Shadow Persons) is important in the APDS-PROMIS System.

Figure 2

PERSON CHARACTERISTICS ARRAY



Pay-offs Associated with Personnel Assignments

Next, it is necessary to determine some indication of effectiveness or pay-offs to the Air Force of assigning a particular person to a particular job. It is desired to find a way to combine different information related to pay-off or value into a single composite indicator. Information from management policy, from operations analysis studies, and human resources research must be combined to yield an indicator of

pay-off. The attempt to obtain such pay-off measures will be done through Policy Development procedures (Ward, 1977). Policy Development includes the combination of Policy Capturing and Policy Specifying. For Policy Capturing, a group of policy makers are presented performance-related information (technical school grades, job performance reports, or predictions of these variables, etc.) about a sample of persons and jobs. The judges (policy makers) will be asked to state the "pay-off" to the Air Force of this sample of persons associated with these particular jobs. Then, a computer will attempt to capture the policy of the judges by developing a mathematical model for predicting the judged values from the person and job information.

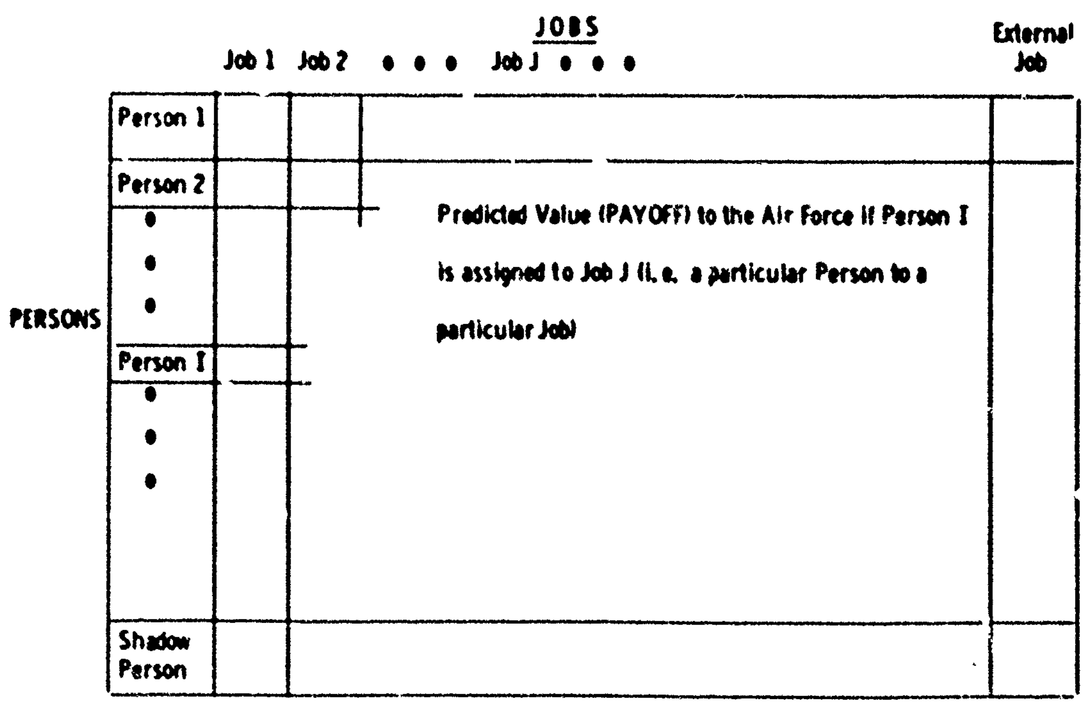
In Policy Specifying, managers express their "pay-off" to the Air Force of various person-job combinations through statements about general constraints that the mathematical model should have. When these constraints are imposed, a model evolves which will produce pay-off values consistent with the specified policy guidelines.

When appropriate, Policy-Specifying and Policy-Capturing can be combined to yield a mathematical model for estimating the value to the Air Force of any person for any Air Force job.

Figure 3, PREDICTED PAY-OFF ARRAY, represents the pay-off values estimated from the mathematical model using the person-job information. The pay-offs associated with the Shadow Person (last row) reflect the values to the Air Force (possibly negative values) of not filling various jobs. The pay-offs associated with the External Job (last column) reflect the values to the Air Force (possibly negative values) of not assigning each person to an Air Force job. In APDS-PROMIS, each applicant is already in an External Job and some applicants are not accepted into Air Force assignments.

Figure 3

PREDICTED PAYOFF ARRAY



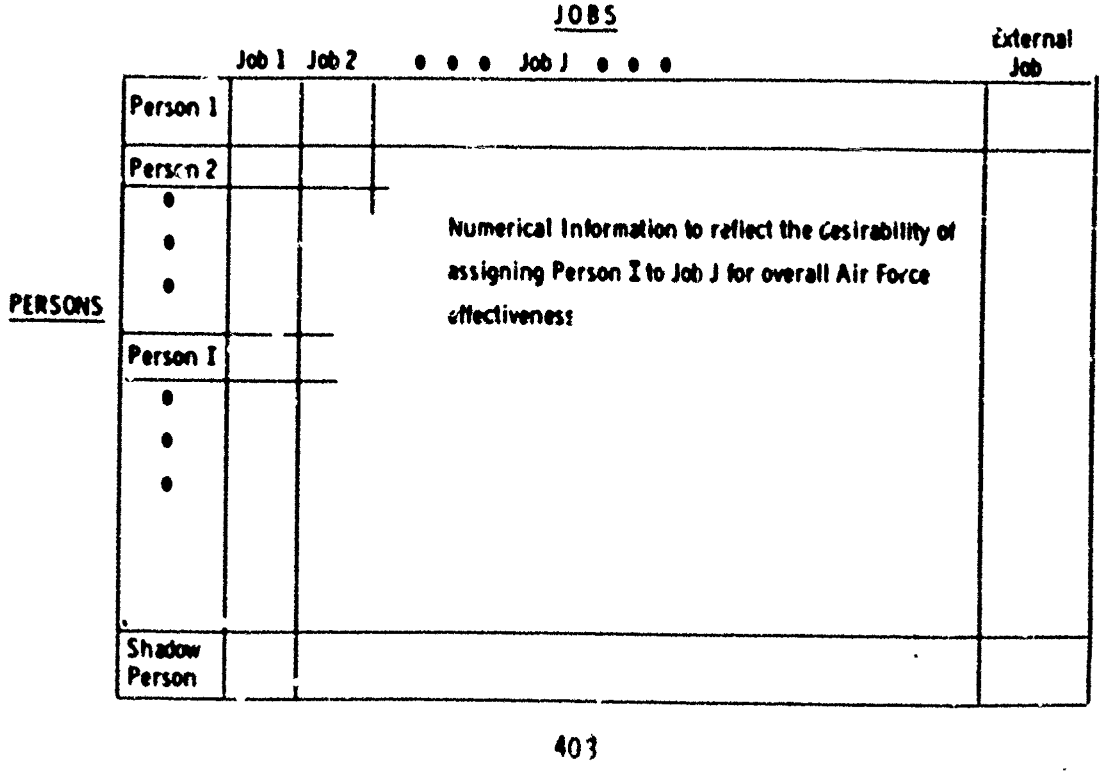
Allocation of Personnel for Overall Air Force Effectiveness

After the elements of the PREDICTED PAY-OFF ARRAY are available, it is necessary to allocate persons to jobs in a way that will tend to maximize overall Air Force effectiveness. The allocation process may not always assign a person to the job for which he has the highest pay-off to the Air Force since many persons must be considered for the job. The attempt is to make assignments that will tend to maximize overall Air Force effectiveness. Figure 4, ALLOCATION ARRAY, contains allocation indicators and represents the information that reflects the desirability for overall Air Force effectiveness of assigning particular persons to particular jobs. This information can reflect the results of an optimal allocation algorithm when appropriate (e.g., Langley's Primal Algorithm (Langley, Kennington, Shetty, 1974)). In this case, the elements of the ALLOCATION ARRAY will contain values of 1 where the assignments result in the maximum overall pay-off and 0 for the non-optimum assignments.

The ALLOCATION ARRAY may also reflect a wide range of numerical values (e.g., Ward's Decision Index (Ward, 1959)) that when used as a basis of assignment will tend toward maximum overall Air Force effectiveness. This approach is appropriate when a sequential-constrained-choice assignment is desired (such as in APDS-PROMIS), the problem is too large for optimum solution, or some of the data required for optimum solution is not available (Ward and Davis, 1963). Both optimum allocation algorithms (for batch assignments) and near-optimum procedures (for sequential-constrained-choice) should be available in a personnel system and used as appropriate.

Figure 4

ALLOCATION ARRAY



EXAMPLE OF PREDICTED PAY-OFF ARRAY AND ALLOCATION ARRAY. Figure 5 illustrates the difference between the PREDICTED PAY-OFF ARRAY and the ALLOCATION ARRAY. The elements of the allocation array reflect that assignment of person 1 to job 3 (allocation index = 14.0), person 2 to job 1 (allocation index = 14.0), and person 3 to job 2 (allocation index = 13.5) will maximize the sum of pay-off values (6 + 5 + 4 = 15). It is interesting to observe that an optimum allocation algorithm would produce an allocation array with values of 1 in the place of the index values 14.0 (Person 1, Job 3), 14.0 (Person 2, Job 1), 13.5 (Person 3, Job 2) to reflect the optimum assignments and 0 in the other 6 locations. However, the values that are now in the array provide for alternative assignments that maintain near optimality. This is operationally important in a system that provides for choice in either a sequential or batch assignment system. A person can be allowed to choose from jobs which have high allocation index values and thereby maintain high overall Air Force effectiveness. For example, if person number 1 were allowed to choose either job 2 or 3 - and he chose job 2 (second highest allocation index) then a pay-off sum of 13 would be possible. (Either 7 + 5 + 1 = 13 or 7 + 0 + 6 = 13).

Figure 5

EXAMPLE OF PREDICTED PAYOFF ARRAY AND ALLOCATION ARRAY

PREDICTED PAYOFF ARRAY

JOBS

PERSONS	Job 1	Job 2	Job 3
Person 1	8	7	6
Person 2	5	1	0
Person 3	6	4	1

ALLOCATION ARRAY

JOBS

PERSONS	Job 1	Job 2	Job 3
Person 1	12.0	13.0	14.0
Person 2	14.0	11.5	12.5
Person 3	13.0	13.5	11.5

The higher numbers in the Allocation Array reflect the desirability of assignments for overall effectiveness of the Air Force

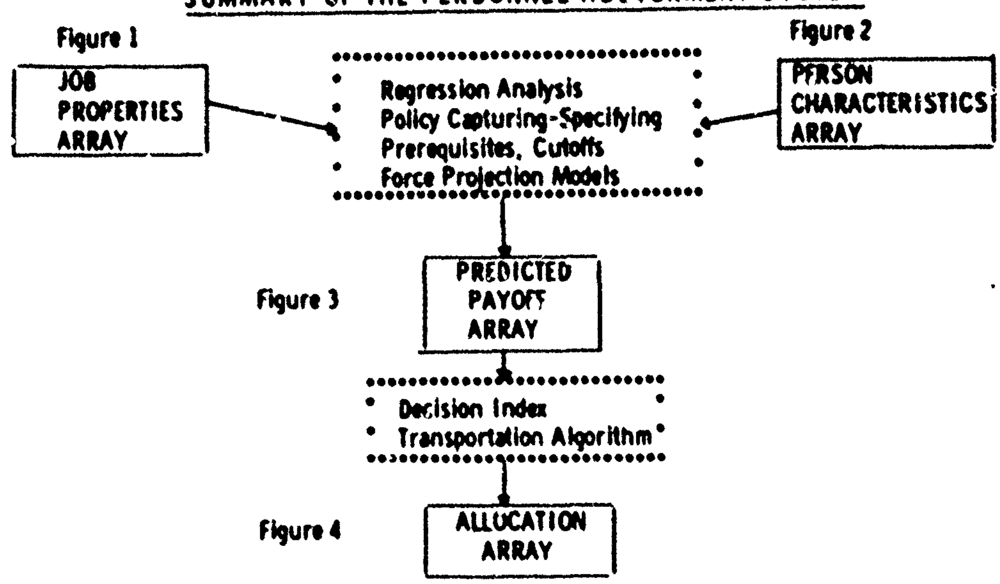
(Overall Effectiveness = 6 + 5 + 4 = 15)
When Highest
Allocation Indexes
Are Used

Summary of the Personnel Assignment System

Figure 6 summarizes the basic features of the personnel assignment system. Information about jobs (Figure 1) and people (Figure 2) are mixed to generate a pay-off (or value) of each potential person-job assignment (Figure 3). From the pay-off array an allocation array (Figure 4) is produced to indicate the appropriateness of each potential assignment for overall Air Force effectiveness.

Figure 6

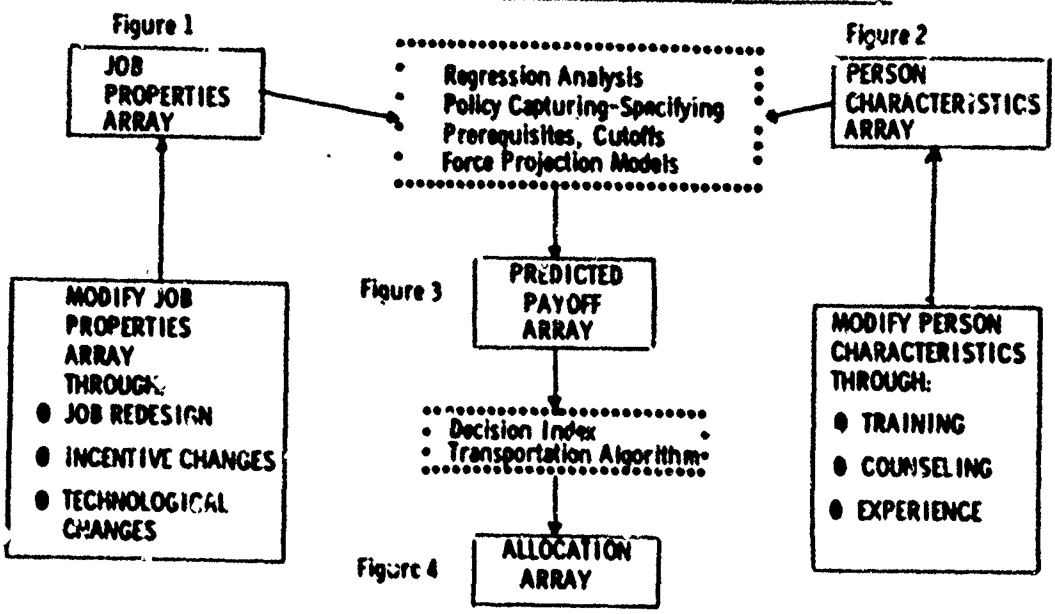
SUMMARY OF THE PERSONNEL ASSIGNMENT SYSTEM



While Figure 6 summarizes the personnel assignment system which considers personnel and jobs as they exist - Figure 7 represents the modification of job properties and modification of person characteristics so that the pay-off array can be improved. Continued personnel training, occupational re-design and organizational improvement can bring about desired changes in personnel and jobs.

Figure 7

A VIEW OF PERSONNEL ASSIGNMENTS
INCLUDING TRAINING AND JOB MODIFICATION



The generalized, modifiable personnel assignment system described above combined with changes in jobs and people characteristics should result in improved Air Force effectiveness. The following section describes the application of these ideas to APDS-PROMIS.

IV. ADVANCED PERSONNEL DATA SYSTEM PROCUREMENT MANAGEMENT INFORMATION SYSTEM (APDS-PROMIS)

The Air Force Recruiting Service lists the characteristics of APDS-PROMIS shown below.

WHAT IS APDS-PROMIS?

- Real-time computer system to replace telephone link
- Job counselling transferred to AFEES processing team
- Computerized preenlistment job classification (P/J match)
- Recruiting objectives for 210 days
- Improved requirement accounting
- Reduced manual reporting
- More professional recruiting image

The following special features were considered in the design of the system:

SPECIAL FEATURES OF

PERSON-JOB MATCH FOR PROMIS ENHANCEMENT

- Sequential consideration of persons to be assigned
- Future accessions are unknown
- List of opportunities must be provided
- Opportunities must be immediately available

SPECIAL FEATURES

FOR ACCEPTANCE AND MAINTENANCE

- Pay-off functions easy to define and modify
- Effects of modifications are easily visible on opportunities list
- * Provide capability through which human resources research findings can affect and improve individual personnel assignments

Opportunity

The major component of PROMIS is the OPPORTUNITY command. The following events provide the ordered list of jobs from which an applicant may choose:

OPPORTUNITY

Person/Job Match

- Input applicant aptitude, physical & preference data
- Test qualification for jobs
- Test availability of jobs
- Compute 'worth' (appropriateness) value for each job
- Maximize total worth to Air Force and individual
- Provide list of most appropriate jobs
  - GTEP
  - Open enlistment
- Offer option to reserve job from list

Predicted Pay-off Values. As indicated above an essential step is the creation of a pay-off array. There are five components contributing to the pay-off values.

CREATING PREDICTED PAY-OFF OF A PERSON-JOB COMBINATION USING POLICY SPECIFYING

* * Person-Aptitude and Job Difficulty (The A-D Component)
* Technical training success
* Aptitude area preferences
* Rate of job fill
* Minority job fill

Aptitude Potential and Job Difficulty. Research findings and experienced personnel people have indicated that interacting a person's aptitude with the job's aptitude requirements so that the most talented people are assigned to the most demanding jobs will reduce training costs, increase job satisfaction and productivity, and improve personnel retainability. This concept has been implemented through the A-D (Aptitude-Difficulty) component.

APTITUDE POTENTIAL AND JOB DIFFICULTY

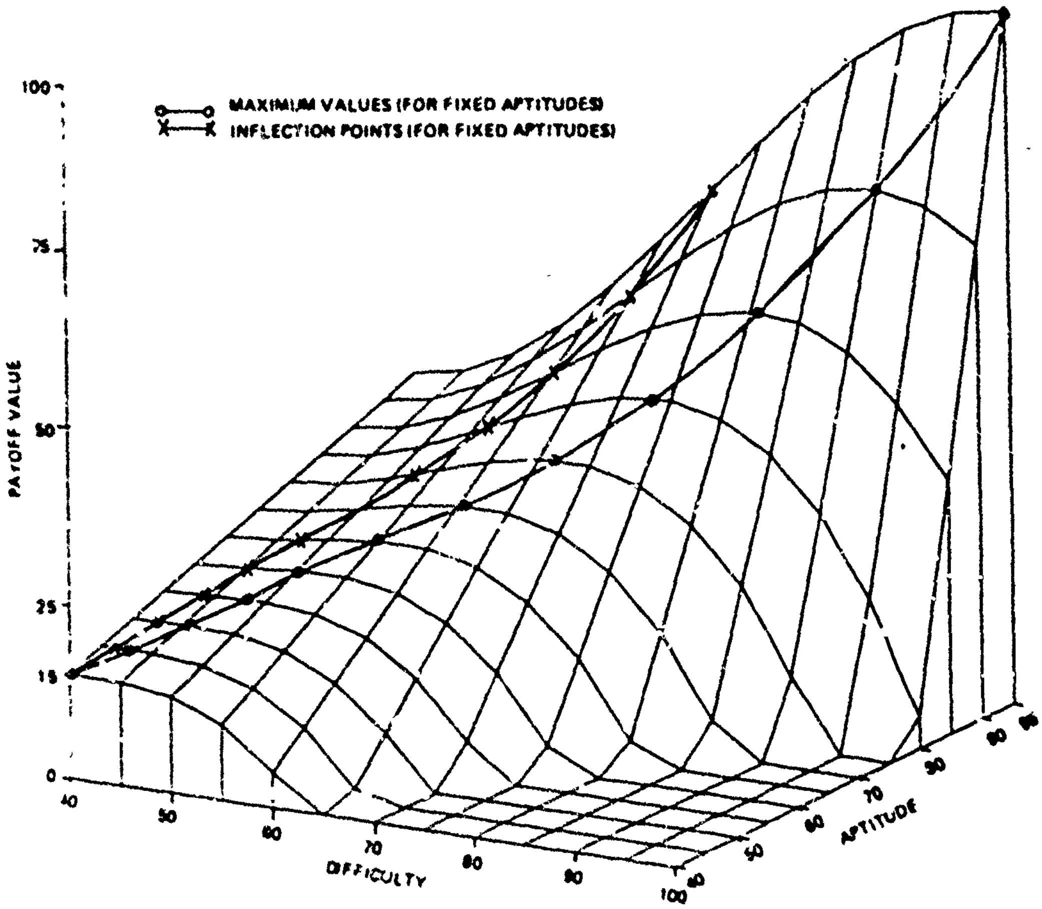
$Y = f(A, D)$

where

A = Aptitude for particular job

D = Relative difficulty of particular job

A three-dimensional view of this component is shown in the following figure.



Pay-off Function of Aptitude and Difficulty

This figure indicates that for a low difficulty job--for example, D = 40--there is a slight increase in pay-off as aptitude increases; however, for a higher difficulty job--for example D = 60--the increase in pay-off is more rapid. Also, notice that for a low aptitude person --for example Aptitude = 40--the highest pay-off is on a low difficulty job, with the pay-off decreasing rapidly as difficulty increases. And for higher aptitude persons the best pay-off is on higher difficulty jobs. A person will have maximum pay-off when his aptitude closely matches the job requirements. And higher aptitudes matched to more difficult jobs are more valuable than lower aptitudes matched to less demanding jobs.

At the present time, only that part of the function to the left (or higher side) of the ridge is getting any use because existing ineligibility rules do not allow applicants who have aptitudes below a certain cut-off to be considered for a job--i.e., the worth below the cut-off is

negative infinity! However, if policy makers allow applicants to become eligible for jobs slightly below existing cut-off scores the pay-off function is available for use. Slight lowering of cut-off rules would allow greater flexibility for making personnel assignments which should result in better use of manpower.

Technical Training Success. The second component is technical training success. This function involves predicted technical school success from aptitude tests, high school courses taken, the particular technical school, and high school graduation status.

TECHNICAL TRAINING SUCCESS

Y = f (AQE, AFQT, HS courses, Tech Schools)

Aptitude Area Reference. Each applicant expresses a relative preference weighting for the four areas -- Mechanical, Administrative, General, and Electronics. These preferences are considered in the pay-off function.

APTITUDE AREA PREFERENCES

Y = f (M, A, G, E preferences)

where

M = Mechanical AI

A = Administrative AI

G = General AI

E = Electronics AI

This component may be replaced in the future by the Vocational Interest Career Examination (VOICE).

Job Fill Rate. This dynamic feedback component is of extreme importance to recruiting service. It reflects interaction between the percentage of jobs sold, amount of time since job was released, and a priority associated with each job. As each job is reserved, and as time changes, this component is modified to change emphasis on jobs that are ahead or behind a desired rate of fill.

JOB FILL RATE

$Y = f(P_j, T, K)$

where

P_j = Percentage of Jobs sold

T = Amount of time since job release

K = Job fill priority

This job fill rate component is in the process of being modified to reflect the actual number of unfilled jobs interacting with the other three job properties -- percentage fill, time, and priority.

Minority Job Fill

This dynamic component is continuously adjusted to maintain a specified minority balance across jobs.

MINORITY JOB FILL

$$Y = f(P_m, G)$$

where

P_m = Percentage of jobs filled by minorities

G = Desired minority job fill goal

Maximizing Overall Air Force Effectiveness. PROMIS requires presentation of an ordered list of jobs from which applicants may choose. An Allocation Index is computed that reflects the desirability for overall Air Force effectiveness of assigning the applicant to each job on the list. An Allocation Index called the Optimality Indicator is used as the basis of ordering. This index is based on the Decision Index (Ward, (1959)) described above.

ASSIGNMENT OF PERSONNEL TO MAXIMIZE OVERALL AIR FORCE EFFECTIVENESS

Decision index used as the allocation index for ordering the opportunities list

V. PLANNED IMPROVEMENTS TO PROMIS

The evolutionary capability of the system allows for incorporating modifications as required. Planned improvements are shown below.

PLANNED IMPROVEMENTS

- Modify fill-rate component to reflect actual number of jobs unsold
- Combine the aptitude-difficulty component interactively with the fill-rate component to reflect policy in which the importance of fill-rate is different for different levels of the aptitude-difficulty component
- Combine attrition prediction information with training costs into the pay-off function to direct good risks to more expensive training and poor risks to less expensive training
- Introduce results from the Vocational Interest Career Examination (VOICE) into the pay-off function to improve job satisfaction and personnel retainability
- Consider interaction of the aptitude-difficulty component with the VOICE (interest) component

VI. APPLICABILITY TO OTHER PERSONNEL SYSTEMS

The concepts above can be applied to any personnel system that would like to match person characteristics with job properties and produce either an ordered list of job opportunities from which an applicant may choose (as in APDS-PROMIS) or an ordered list of applicants from which a job manager may choose (as when a job must be filled). The airmen post-enlistment assignment system, now being developed, should be applicable to a wide variety of personnel sub-systems--airmen, officers, and civilians.

APPLICABILITY TO OTHER PERSONNEL SYSTEMS

- Air Force enlisted re-assignments
- Officers assignments
- AF civilians
- Others

VII. SUMMARY OBSERVATIONS

A mechanism is evolving through which human resources research findings can directly affect and improve individual personnel assignments. System flexibility provides for modification and introduction of new components to insure continued acceptance and improvement. The approach has general applicability to personnel systems that can identify information about persons and jobs and specify a pay-off generating policy.

Implementation of this approach has led to identification of areas of human resources research that will contribute significantly to improved systems performance.

RESEARCH AREAS OF POTENTIAL VALUE

- SEARCH FOR PERSON CHARACTERISTICS AND JOB PROPERTIES THAT INTERACT IN PREDICTION OF PAY-OFF VALUES
- DEVELOP NEW METHODS FOR SPECIFYING THE PAY-OFF VALUES ASSOCIATED WITH PERSON-JOB ASSIGNMENTS
- STUDY THE USE OF ALLOCATION INDEXES NOT ONLY AS AN ORDERING VALUE FOR OPPORTUNITY LISTS, BUT AS A SUPPLEMENT TO APTITUDE INDEXES NOW IN USE

REFERENCES

Langley, R.W., Kennington, J., & Shetty, C.M. Efficient computational devices for the capacitated transportation problem. Naval Research Logistics Quarterly, Office of Naval Research, December, 1974, v. 21, No. 4, 637-647.

Pina, M.,Jr. & Stifle, J.L. Person-job match computer-based research system. AFHRL-TR-77- , Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, in press.

Ward, J.H.,Jr. Use of a decision index in assigning air force personnel. WADC-TN-59-38, AD-214 600. Lackland AFB, TX.: Personnel Laboratory, Wright Air Development Center, Air Research and Development Command, April 1959.

Ward, J.H.,Jr. & Davis, K. Teaching a digital computer to assist in making decisions. PRL-TDR-63-16, AD-407 322. Lackland AFB, TX.: 6570th Personnel Research Laboratory, Aerospace Medical Division, June 1963.

Ward, J.H.,Jr., & Haltman, H.P. Computer-based enlistment quota reservation system using the general data management system 2000: programming and implementation details. AFHRL-TR-75-71, AD-021 340. Lackland AFB, TX.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, December 1975.

Ward, J.H.,Jr. Creating mathematical models of judgment processes: From policy-capturing to policy-specifying. AFHRL-TR-77-47. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, August 1977.

QUANTIFICATION OF PERCEPTUAL DIFFERENCES BETWEEN EVALUATORS AND EVALUATED UNITS

A Paper Based on an

Evaluation of the

11th Special Forces Group--Airborne

by

Earl W. Rubright

Presented to

Military Test Association
San Antonio, Texas

19 Oct. 1977

On October 18, 1976, the 80th Maneuver Training Command received the requirement to produce the documents required to evaluate the combat readiness of the 11th Special Forces Group--Airborne, a group targeted against Eastern Europe. The actual adaptation of Army Training Evaluation Program 31-101 (ARTEP 31-101), the primary evaluation tool, was the responsibility of the Infantry Team in general and the author in particular. The author realized that his biases were based on his combat experience in southeast Asia and that these biases had been reinforced by the three years he spent in a Special Forces unit targeted against Asia. Based on this the author recommended that the production of the ARTEP be assigned to an individual with extensive European Special Forces experience or an attempt be made at quantifying the perceptions of those who had European Special Forces experience. The decision was made to quantify the perceptions of those involved in the evaluation. This decision was based on the assumption that underlies decentralized training, i.e., that the Commanding Officers will train their units in accordance with their perceptions of the combat requirements of the area they are targeted against (in the case of Special Forces) and will use the ARTEP as a guide.

The ARTEP was dismembered and each combat requirement that it contained became part of a pool of combat requirements that were used for a Q-sort. The Q-sort was administered to Special Forces qualified persons who had extensive European experience. Those specific combat requirements of the ARTEP that were selected as being important to/for a Special Forces Group targeted against Europe became the stimuli for a pair comparisons test.

The pair comparison was administered to the Deputy Commanding

Officer (DCO), Operations (S-3), Intelligence (S-2), and Supply Officers (S-4), plus the 2nd Battalion, 3rd Battalion and Support Battalion Commanding Officers. Three Operational Detachment Commanding Officers (ODA) of the 11th Special Forces Group--Airborne (SFGA) were also included.

The question then became, "Are the perceptions as measured by the Pair Comparison a true reflection of the relative importance of the combat requirements of Eastern Europe?" The answer to this question was needed to ensure that an evaluation stressing the combat requirements of greatest relative import was a true reflection of the actual combat requirements of Eastern Europe. The answer to this question was gained by administering the Pair Comparison to a similar population in the 10th Special Forces, a highly regarded active Army unit targeted against the same area. The administration of the Pair Comparison to the 10th Special Forces served two purposes. First, it gave an indicator of the accuracy of the 11th SFGA's perceptions of the relative import of the combat requirements of Eastern Europe. Second, since the 10th Special Forces was furnishing the actual on-the-ground evaluators, an indicator could be gained as to whether differences in evaluator-evaluatee perceptions would be likely to skew the evaluation.

The results of both administrations of the Pair Comparisons indicated that the focus of the evaluation and training should be one the training and use of the guerrillas. Because of the close coordination required to develop and sustain a guerrilla force, the decision was made to develop three situations. They consisted of:

1. An operation in which the selected operational capabilities of the Operational Detachment are measured (ARTEP 31-101).

2. A situation that stressed the importance of the guerrilla through intelligence play.

3. A brief situation that stressed S-2 (Intelligence) and S-3 (Operations) coordination.

The use of guerrillas can be divided into two areas:

1. Operations support, which ARTEP 31-101 encompasses.

2. Intelligence support at both a tactical and strategic level.

The lack of intelligence requirements in ARTEP 31-101 required the generation of a situation in which intelligence reports would be transmitted both up and down the chain of command. This was accomplished by interrelating intelligence reports from different levels. An example follows.

SOCGULF (Evaluator control HQ)

Ft. Rucker, AL — SFOB Special Forces Operations Base (11th SFGA HQ)

Ft. Stewart GA — ACB — 2nd Bn 11th SFGA

Ft. Benning, GA — ACB — 3rd Bn 11th SFGA

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX FEBA XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

ODA ODA ODA ODA

G G G G G G G G

ODA ODA ODA ODA

G G G G G G G G

In the play of the problem, the fact that the 12th Combined Arms Army (Soviet) was moving two tank divisions forward to reinforce was the result of many indicators from each level. That tracked vehicles were moving forward was based on Side Looking Airborne Radar (SLAR) reports that were received at Special Forces Operations Base (SFOB). That the tracked vehicles

were predominately tanks was based on guerrilla agent reports given to the Operational Detachment (ODA). That the tanks were from at least two different units was based on the correct interpretation of photos taken by the guerrillas, passed to the Operational Detachment, and exfiltrated from the operational area and returned to the Advanced Control Base (ACB) level and eventually to the Special Forces Operations Base level.

For the full story to emerge, the requirement was for each level to synthesize the information it received and pass it on to both superior and subordinate headquarters. The interrelating of intelligence reports across levels prevented any single headquarters from independently piecing together the whole story.

This technique was employed to portray:

1. The forward deployment of two tank divisions.

2. Combat tailoring of enemy units to counter the increased insurgent threat being developed by the Operational Detachments as they trained the guerrillas (guerrillas were actually high school ROTC cadets).

3. New troop deployments indicating a build up to halt the impending friendly conventional offensive.

Each Operational Detachment received the same intelligence reports from the guerrillas. This allowed the Battalion S-2 (Intelligence Officer) to assess the relative efficiency of each Operational Detachment to process and forward the information received from the guerrillas. Where the photos of enemy equipment given to the Operational Detachment by the guerrillas could not be interpreted and forwarded as messages, the imagery itself was to be exfiltrated via the Fulton Recovery system.

Similar procedures were required of the Special Forces Operations Base when receiving information from SOCGULF (Evaluator Control Head-

quarters), Advanced Control Bases, or Operational Detachments. The requirement was to synthesize the information with information already on file, draw conclusions and forward the intelligence derived by either radio message if it was of an immediate/critical nature or by a daily publication called an INTSUM (Intelligence Summary). Photos that could not be interpreted at a lower level were the responsibility of the Imagery Interpretation section of the Group Military Intelligence Detachment.

A third situation was based on the 11th SFGA resupply request message format. An aerial resupply operation calls for the approving authority of that mission to integrate all information available to ensure not only the success of the mission but the survival of the aircraft. This is accomplished by notifying the Air Force element that is to fly the mission of any natural or manmade hazards that are associated with a particular drop zone. The notification can take several forms--the most common of which is requiring the aircraft to follow a specific track or azmuth when crossing over the Drop Zone.

The performance of the 11th SFGA could then be judged on three criteria:

1. The extent to which they adequately performed tasks required in ARTEP 31-101.

2. The extent to which the Operational Detachments, Advanced Control Bases, and Special Forces Operations Base were able to alert SOCGULF to the:

a. forward deployment of two tank divisions

b. combat tailoring to meet the guerrilla threat

c. new troop deployments.

3. The ability of the S-2 and S-3 sections to coordinate their activities.

The formal presentation of the evaluation based on ARTEP 31-101 will not be made until October, 1977. It does appear that with the exception of communication that the Operational Detachments did reasonably well on the operational portions of the operation. The result of the intelligence portion (#2) was less encouraging. The Operational Detachments failed to communicate the information derived from the guerrillas to the Advanced Control Bases and Special Forces Operations Base. This failure cannot be blamed on communication difficulties alone. The Fulton recovery extracted no intelligence summaries, no raw intelligence reports, no agent reports and no imagery. Similarly, there was no indication that Special Forces Operations Base or Advanced Control Base INTSUMS were received by the Operational Detachments. One particularly disheartening aspect of the intelligence portion of this operation was that the Imagery Interpretation section of the 11th Special Forces' Military Intelligence Detachment was unable, because of lack of photo keys, to interpret over 85% of the imagery that was administratively provided to them. This was the same imagery given to each Operational Detachment by the guerrillas, all of which was organic to the Soviet Motorized Rifle Division or Combined Arms Army. More disheartening was the fact that the Imagery Interpretors were unable to interpret major end items, i.e., tanks, self-propelled artillery, or to tell the author where, by number and location, in the Soviet system these weapons would be found.

The third situation generated tested the ability of the S-2 and S-3 sections of the 11th Special Forces to coordinate or tie together their activities for the conduct of an aerial resupply mission. To this end intelligence reports positioning Surface-To-Air Missile (SAM) complexes just out of range of several resupply Drop Zones (DZ) was forwarded from

SOCGULF to Special Forces Operations Base. The requirement was for each message/Drop Zone request sent to the US Air Force to have a required magnetic azmuth that would route the resupply aircraft out of SAM range incorporated in the request. This would have required that the S-2 receive the message, mark on a map the positions of SAM and determine range fans for each SAM complex. Coordination would then have to be made with the S-3 to ensure that the Drop Zone request included the required azmuth to protect the USAF resupply aircraft. None of the Drop Zone requests forwarded to the USAF included a required azmuth. The inability of the 11th SFGA to tie the operations and intelligence sides of this operation together in this instance indicates serious training deficiencies that go beyond the scope of this evaluation.

Of the three situations generated for the evaluation of the 11th, only one was based on ARTEP 31-101. A conclusion noting that the 11th SFGA was testbound would, however, be premature. More likely, it is the carry over effect of past Army Training Tests which were almost totally operations oriented.

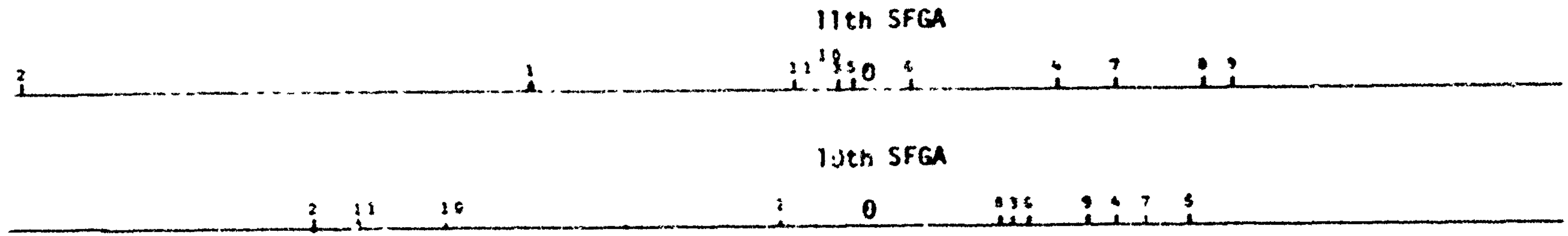
The use of the perceptions of the 11th SFGA to focus the evaluation was, as stated earlier, based on the premise that the commanders will train for what they consider important. Unfortunately, the results of the Pair Comparisons test, when contrasted to the training conducted by the 11th SFGA, highlighted additional problems. The Pair Comparison results of both the 10th and 11th Special Forces place Airborne Operations at the very bottom of the priority list and the ability to effectively train a guerrilla force at the very top of the priority list. Yet, during Annual Training the 11th Special Forces spent approximately $300,000 on air operations and none on language training--a capability which,

historically, has dictated the success or failure of an unconventional warfare effort. The almost perfect inverse relationship between perceived importance of combat requirements and resource allocation for combat requirements casts doubt on the authenticity of the mission statement for Reserve Special Forces. The discrepancy between mission and resources allocated by regulation leaves the command of the 11th Special Forces in a position of being unable to address those aspects of Special Forces operations that are considered essential (by both the 11th Special Forces and 10th Special Forces) to the success of combat operations.

COMBAT REQUIREMENTS AS PERCEIVED BY THE 10th

and 11th SPECIAL FORCES GROUP-AIRBORNE

10th SFGA Value	11th SFGA Value	Stimuli
-.17	-.56	1. Infiltrate by parachute (static line)
-1.01	-1.62	2. Infiltrate by parachute (HALO)
+.32	-.05	3. Assemble detachment personnel and account for equipment
+.49	+.35	4. Contact resistance force reception committee
+.61	-.02	5. Sterilize infiltration site
+.30	+.08	6. Begin movement to designate base
+.53	+.48	7. Begin area assessment and contact SFOB
+.28	+.65	8. Organize guerrilla force with command and staff
+.43	+.69	9. Develop a training program for the guerrilla force
-.77	-.05	10. Select and report LZs/DZs
-.99	-.11	11. Secure, mark and operate LZs/DZs



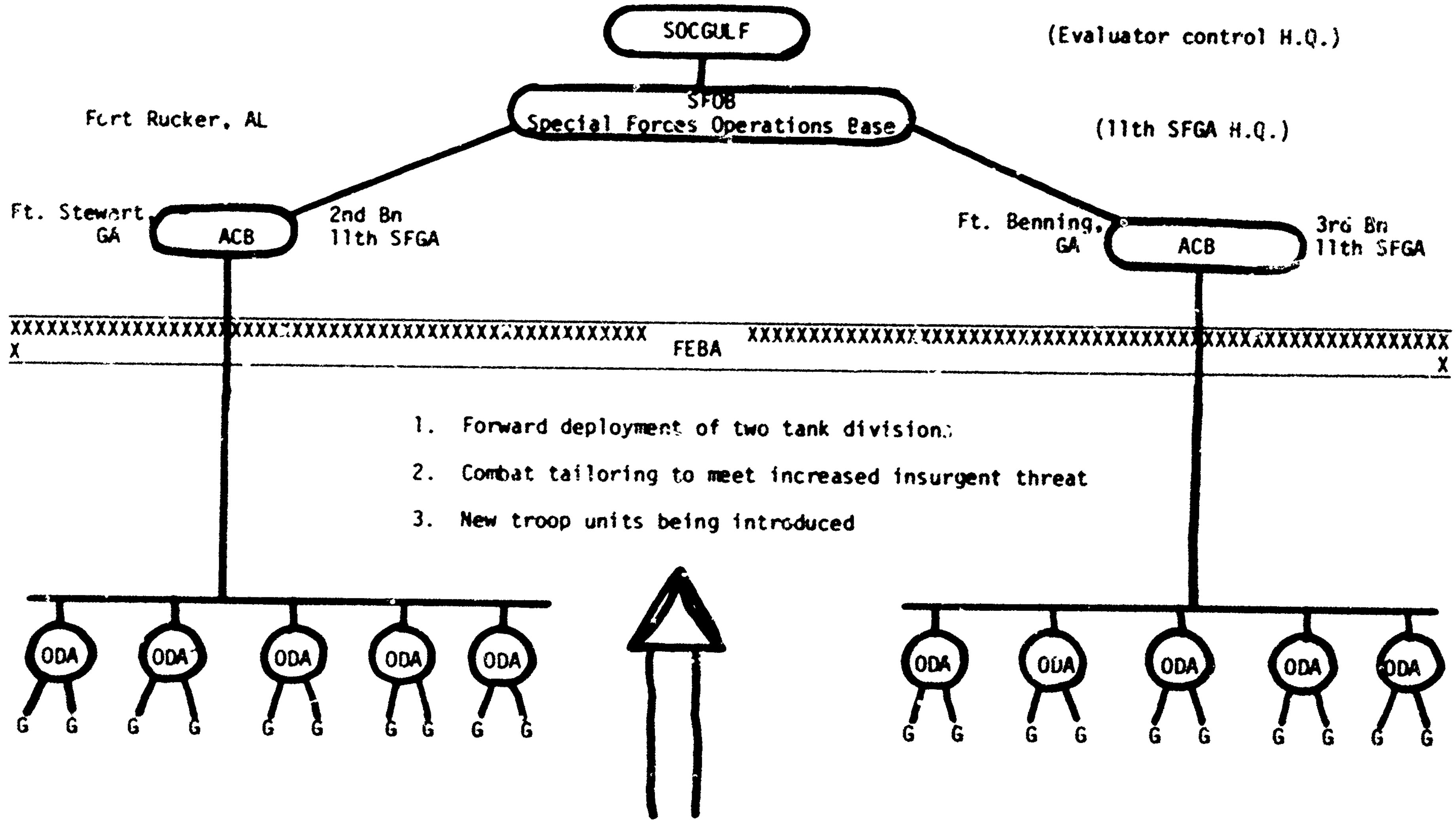
10th SF	11th SF		
-.17	-.66	1.	Infiltrate by parachute (static line)
-1.01	-1.62	2.	Infiltrate by parachute (high altitude, low opening [HALO])
.32	-.05	3.	Assemble detachment personnel in accordance with established procedures and account for accompanying equipment.
.49	.35	4.	Contact the resistance force reception committee in accordance with established procedures (when a reception committee is at the infiltration site).
.61	-.02	5.	Sterilize the infiltration site to preclude its being identified as such.
.30	.08	6.	Quickly begin movement to designate base, contact, or safe areas, using tactical march techniques.
.53	.48	7.	Upon arrival in the base area, immediately begin area assessment and contact SFOB to indicate infiltration is completed.
.28	.65	8.	Organize the guerrilla force into recognizable military units with a viable command and staff structure.
.43	.69	9.	Develop a training program that will enhance the operational capabilities of the guerrilla force.
-.77	-.05	10.	Select and report landing zones and drop zones (LZs/DZs) that meet aircraft and resupply requirements.
-.99	-.11	11.	Secure, mark and operate LZs and DZs.

COMPONENTS OF 11TH SPECIAL FORCES EVALUATION

1. A measure of selected operational requirements, i.e., ARTEP 31-101

2. A situation that stressed Intelligence play.

3. A situation that stressed Intelligence and Operations coordination


SOCGULF
(Evaluator control H.Q.)
SFOB
Special Forces Operations Base
Fort Rucker, AL
(11th SFGA H.Q.)
Ft. Stewart, GA
ACB
2nd Bn
11th SFGA
Ft. Benning, GA
ACB
3rd Bn
11th SFGA
FEBA
1. Forward deployment of two tank divisions
2. Combat tailoring to meet increased insurgent threat
3. New troop units being introduced
ODA
G

CRITERIA FOR EVALUATION

1. Performance on Operational Tasks of ARTEP 31-101

2. Ability to alert SOCGULF to:

   a. forward deployment of two tank divisions

   b. combat tailoring to meet guerrilla threat

   c. new troop deployments

3. Ability of S-2 and S-3 sections to coordinate their activities

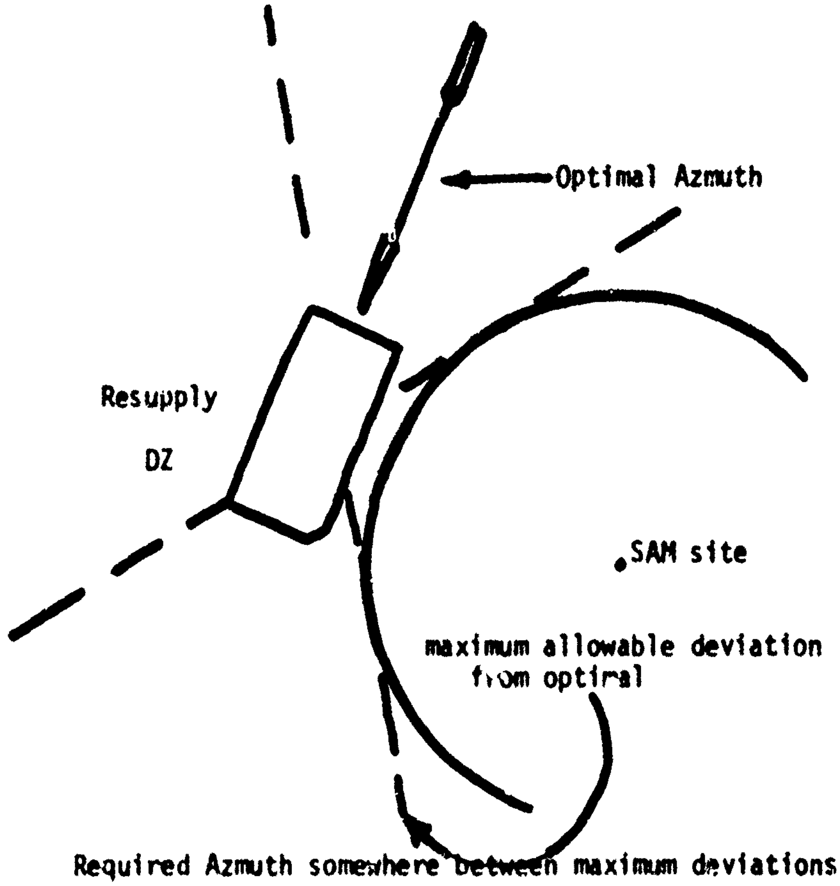
IMAGERY INTERPRETATION RESULTS

1. Division Organic Combat Equipment = 15% identified

2. Division Organic Electronic Warfare Equipment = 0% identified

3. Combined Arms Army Organic Electronic Warfare/ Communication Equipment = 0% identified

S-2/S-3 COORDINATION



RESULTS OF S-2/S-3

COORDINATION REQUIREMENT

Number of Possible DZ Coordinations	Number of DZs Coordinated
12	0

EVALUATION CRITERIA

? ARTEP 31-101

X Intelligence requirement

X Coordination requirement

Bibliography

Ferguson, George A. Statistical Analysis in Psychology and Education. New York: McGraw-Hill Book Company, 1959.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill Book Company, 1954.

Nie, Norman H. and others. Statistical Package for the Social Sciences. New York: McGraw-Hill Book Company, 1970.

Siegel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company, 1956.

Struening, E. L. and Guttentag, M. Handbook of Evaluation Research, Volume II. Beverly Hills, California: Sage Publications, 1975.

U.S. Army Field Manual 31-20 Special Forces Operational Techniques.

U.S. Army Field Manual 31-20A Special Forces Operations.

U.S. Army Field Manual 31-21 Special Forces Operations.

U.S. Army Field Manual 31-21A Special Forces Operations.

U.S. Army Field Manual 30-102 Handbook on Aggressor Military Forces

U.S. Army Field Manual 30-5 Combat Intelligence.

U.S. Army Field Manual 30-103 Aggressor Order of Battle.

U.S. Army Field Manual 101-5 Staff Operations.

U.S. Army Field Manual 105-5 Maneuver Control.

U.S. Army Reserve Training and Evaluation Program 31-101 Special Forces Group Airborne

U.S. Army Modified Table of Organization and Equipment 31-310H Special Forces Group Airborne

CRITERION-REFERENCED SYSTEMS APPROACH TO EVALUATION OF COMBAT UNITS

Angelo Mirabella

Army Research Institute for the Behavioral and Social Sciences

System Engineering of training and its subsidiary criterion-referenced measurement have been invaluable tools for increasing the job-relevance of military training and evaluation. These tools have provided an indispensible point of departure and a framework for insuring accountability. However, they have been developed within the context of relatively simple, procedural tasks that are necessary but not always sufficient for describing jobs as performed in working environments. The tools work comfortably for hard, individual skills. But the demurrer often heard is that the soft skills have yet to be attacked successfully with those tools. Those of us who have moved the focus of our evaluation research from the school setting to the combat unit are especially sensitive to this demurrer because we face the added complication of two-sided, tactical, collective behavior.

In order to improve the training and evaluation of such behavior, the Army Research Institute for the Behavioral Sciences has been pursuing research on Tactical Engagement Simulation. In addition, we have been developing a supporting system of evaluation. It is the evaluation system research that I would like to talk about, but for those of you not familiar with our program let be briefly describe the Engagement Simulation Test Bed.

Engagement Simulation currently is a set of techniques of conducting real-time, two-sided free play, tactical exercises at the combined arms reinforced platoon level. One of its key features is a set of objective, casualty assessment methods which allow almost real-time feedback to participants. A rifleman, for example, can fire at a target, and register a hit by calling out a number on the helmet of the opposing infantryman. A tank gunner can similarly register a hit against another tank. Kills are relayed via radio by a controller to a net control station, which in turn radios the target that it is out of action. Suitable pyrotechnics add visual cues and, therefore, realism to the battle[1]. With these and other techniques for artillery and anti-tank weapons, it is possible to measure casualties over time and thereby provide for objective assessment of the outcomes of tactical performance.

Several years ago when Engagement Simulation was developed as a training methodology, its developers felt that the evaluation problem for unit training had been solved. Objective measures of casualties were now available! What else was needed? However, an alternative view was that a great deal else was needed; that the Engagement Simulation

[1]TC 71-5. REALTRAIN: Tactical Training for Combined Arms Elements. U.S. Army Armor School/U.S. Army Research Institute, January, 1975.

test bed had opened up a Pandora's box with respect to the measurement, and interpretation of unit combat performance. This alternative view argued for a system of evaluation which included at least concern for process measurements, and a scheme for uncovering the patterns and relationship among these two sets of measures, taken continually through a training exercise. Such a system included some other major features which I would like to mention.

In fact, I would like to outline what an adequate system of evaluation might look like for the Engagement Simulation test bed, and mention some of our research experiences with various parts of the System (Fig 1). If we were to proceed logically and efficiently through the development of an Engagement Simulation relevant evaluation system we would begin with the development of a model or model(s) to define:

- measurement concepts
- data processing concepts
- data interpretation concepts

which are consistent with the purposes for measuring and assessing performance in the first place. It's at this point that evaluation aims and philosophical biases can be put on the line. If the major purpose of evaluation is diagnostic feedback in support of a training system, that purpose can be made explicit and the rest of the system designed accordingly. This last statement may seem obvious and self evident, but in practice, it may not be so obvious. One of the philosophical problems with the Army Training and Evaluation Program (ARTEP)[2], is that it does not distinguish adequately between evaluation for training diagnosis, and evaluation for accountability. A result has been that many commanders regard ARTEP as a report card in spite of TRADOC's guidance to the contrary. This observation which came out of a current ARI study suggests at least one fundamental problem with ARTEP as a training model[3].

The next step in system development would be to define the data requirements and data processing methods that are needed to fit the model or model(s) constructed in Step 1. If, in Step 1, for example, you decided that information about patterns of tactical movement is useful for diagnostic purposes, that would suggest a need to know what fire elements are where, when. You would need to go further and decide how much information on position location is needed and how accurate it needs to be.

Now you are faced with Step 3 which requires that you define the methods for collecting the data identified in Step 2. If you are not yet familiar with the realities of collecting objective performance data

---

[2]ARTEP 71-2. Army Training and Evaluation Program for Mechanized Infantry/Tank Task Force, June, 1977.

[3]Human Sciences Research, Inc. Interim Report (Revised), Improved Army Training and Evaluation Program (ARTEP) Methods for Unit Evaluation, 21 October, 1977.

FIG 1: ENGAGEMENT SIMULATION EVALUATION SYSTEM

STEPS IN DEVELOPMENT OF SYSTEM:

1. DEVELOP A MODEL OR MODEL(s) TO DEFINE:
   - MEASUREMENT CONCEPTS
   - DATA PROCESSING CONCEPTS
   - DATA INTERPRETATION CONCEPTS
2. DEFINE DATA REQUIREMENTS AND DEVELOP PROCESSING METHODS
3. DEFINE AND DEVELOP DATA COLLECTION METHODS.
4. DEFINE, DEVELOP, PERFORMANCE BENCHMARKING TECHNIQUES, I.E., STANDARDS.

under field operational conditions, you would soon learn about them, at this stage of system development.

Last, but certainly not least, is Step 4 which is to define the performance benchmarks or standards which make your system criterion-referenced. This is probably the most difficult of all. It has been sidestepped to a large extent by ARTEP through the use of expressions like: "Casualties shall not be excessive", with the definition of the benchmark being left to the evaluation team. ARTEP has also sidestepped the criterion issue by using mostly procedural standards, which are at a more global level than those in the old Army Training Tests, but which are still procedures-based.

I would like now to review some of the progress that ARI has made in contributing to such a system.

Modeling. As part of a long-term effort to Validate Engagement Simulation Training, new experimental versions of ARTEP are being produced. These are being produced specifically for some developmental tests to be run at Ft Carson in January of next year. Accordingly, they are being designed for reinforced platoon missions, i.e., for tank platoons with supporting infantry squads and tube launched optically-tracked wire guided missiles (TOWS).

For those of you not familiar with ARTEPs let me show you a typical page, this one from ARTEP 7-45 FOR MECHANIZED INFANTRY AND COMBINED ARMS TASK FORCE. (Fig 2) Look particularly at the column labeled TRAINING/EVALUATION STANDARDS. Phrases like "Coordination...will support...", "must be responsive", "without sustaining excessive casualties", place a substantial responsibility on the evaluator. Now look at a roughly comparable version of an engagement simulation ARTEP: At least two major revisions in ARTEP concept can be seen. (Fig 3)

o The standards and the rating columns have been eliminated. In their places are a performance data and results section. The measures are quantitative: e.g., time, range, casualties.

o Major objectives have been further analyzed into intermediate objectives. For example, the task of eliminating enemy resistance has been analyzed into the weapons systems involved and then further broken down into weapon systems sub-tasks.

What's happened to the standards? This particular model of ARTEP candidly admits that we don't know how to handle the standards problem yet and moves the problem to one side, until scientific progress in this area provides some useful methodology. Emphasis shifts here away from GO/NO GO type of evaluation. Emphasis is placed instead on obtaining a rich, detailed description of the behaviors involved in two sided combat. That emphasis leads to two essential questions:

o What patterns of behavior can we extract from the various performance measures which will have diagnostic value? We have, for example,

FIG 2: TRAINING AND EVALUATION OUTLINE

UNIT: COMPANY/TEAMS

MISSION: DELIBERATE DAYLIGHT ATTACK

TASK	CONDITIONS	TRAINING/EVALUATION STANDARDS
ELIMINATE ENEMY RESISTANCE.	ENEMY ENGAGES LEAD ELEMENTS OF THE TF WHEN THEY ARE WITHIN ANTITANK RANGE OF HIS POSITION (E.G., 1,000 - 2,000 METERS).	A. COORDINATION OF MORTAR, MACHINEGUN, ANTITANK, AND NONORGANIC FIRE SUPPORT WILL SUPPORT THE SCHEME OF MANEUVER (I.E., FIRE MUST BEGIN, BE SHIFTED, AND BE STOPPED AS SPECIFIED IN THE ORDER AND MUST BE RESPONSIVE TO REQUESTS FROM MANEUVER PLATOONS). B. COMPANY/TEAM ELEMENTS USE PROPER FIRE AND MANEUVER TECHNIQUES TO ELIMINATE ENEMY RESISTANCE. OBJECTIVE IS SECURED WITHOUT SUSTAINING EXCESSIVE CASUALTIES AND VEHICLE LOSSES. (EVALUATOR JUDGMENT.)

FIG 3: TRAINING AND EVALUATION OUTLINE

UNIT: COMPANY/TEAM | MISSION: ATTACK

TASK/SUBTASK	CONDITIONS	PERFORMANCE DATA	RESULTS
EXECUTE THE ATTACK (TANKS)	ENEMY FORCE DEFENDS		
CROSS THE LD		TIME TANKS CROSS LD	
TANKS MOVE BY COVERED AND CONCEALED ROUTE TO ATTACK		TANK CASUALTIES DURING MOVEMENT	
TANKS DETECT THE ENEMY AT MAXIMUM RANGE		TIME OF DETECTION DETECTION RANGE	
TANKS ENGAGE THE ENEMY AT MAXIMUM RANGE		ENGAGEMENT RANGE	
TANKS ENGAGE THE ENEMY		CASUALTIES: OPPOSITION OWN KEY PERSONNEL PERSONNEL WEAPONS DISABLED VEHICLES IMMOBILIZED	
EXECUTE THE ATTACK (INFANTRY)			
•			
•			
•			

a particular interest in showing the connection or correlations among tactical movements (i.e., position by time), processes such as first enemy detections, and outcome measures such as casualties inflicted or sustained.

o What performance trade-offs can we identify and measure? A commander may deliberately sacrifice cover and concealment in order to fight more aggresively or move more quickly toward some tactical objective. The significant and diagnostically useful measurement concept would be risk-taking behavior, instead of just cover and concealment.

Again let me ask the question: "What has happened to standards?" We haven't forgotten about them. Until the standards problem is solved, an evaluation system is not criterion referenced. But we have concluded that some imaginative and fresh thinking is required here along with supporting research. The concept which we are currently working on can be described as situation-specific forecasting of the dynamics of an engagement simulation exercise along with various tactical processes and outcomes such as casualties. I'll say a little more about this in a few minutes.

The second step in the evolution of an evaluation system is to define performance data requirements and data processing techniques. The modeling of Step 1 can provide the general guidance for this step. But more specifically our approach has been to identify essential elements of analysis (EEAs) and then to produce measures of effectiveness (MOEs) by phase of combat. This is consistent with the ES ARTEP model. Under Contract to ARI, Human Systems, Inc. of Memphis (HSI) has generated a computer listing of EEAs along with methods for coding, processing, and displaying the results of a computer analysis of a tactical map[4]. Fig 4 shows the initial list which HSI generated. What the data file does, in effect, is to describe and display the tactical movements of two opposing combat teams Alpha and Bravo, involved in an Engagement Simulation exercise. The list indicates which fire elements are in what locations, at what time, what the terrain is like, whether or not there are targets of opportunity and what casualties are resulting from direct and indirect fire.

Currently, these EEAs are being put together in various ways to provide Measures of Effectiveness (MOEs) for each of the phases of reinforced platoon attack mission. This mission will be the basis for a developmental test at Ft Carson in January. The Phases will be:

o Preparation (i.e., Planning)

o Pre-Engagement (i.e., Movement to Contact)

o Engagement (Hostilities)

o Post-Engagement (Post-Attack Security)

---

[4]Hansen, D.N. and Drewfs. Small Unit Data Input Structure and Graphic Support System. Interim Report. Human Systems, Inc., 28 June 1977

FIG 4: EEAs BEING ADDRESSED BY ENGAGEMENT SIMULATION DATA FILE

1. Team Membership
2. Trajectory Force Mix
3. Trajectory Identification Number
4. Trajectory Path Hex Number
5. Elapsed Exercise Time
6. Hex Terrain Type
7. Trajectory Mode
8. Communications
9. Trajectory Plan Versus Actual Hex Position
10. Trajectory Overwatch Placements
11. Trajectory Potential Fire Fan and Coverage Fire Fan
12. Trajectory Movement Rates
13. Target of Opportunity Index
14. Attack Effectiveness (Direct Fire Effects)
15. Incoming Indirect Fire Effects

Some examples of Measures of Effectiveness which we are looking at might concretize what I am saying: Fig 5 shows an MOE for Tactical Formation in the Pre-Engagement Phase. An example from the Engagement Phase is shown in Fig 6. In this case an MOE for enemy detection is illustrated.

The methods for processing data in the HSI/ARI data file are somewhat constrained by their small quantity. Tactical exercises unlike most individual tasks require several hours if not whole days to complete, and are very costly in manpower and supplies. Consequently, data are relatively scarce and do not readily lend themselves to sohpisticated multivariate analysis. Therefore, until a large data base is built up, we will probably not be able to do much beyond cross tabulations of frequency counts. This is the tact which we have been taking so far. But such a tact is consistent with our near term goal which is to build up experience with objective measurement of tactical exercises and to learn which measures are most useful for diagnosis of training deficiency.

Having defined data requirements and even modest data processing approaches, the next step in evolving an evaluation system is to define data collection methods. From a technological point of view collection of position location information is our hottest chestnut. The problem is a very critical one because position location, i.e., tactical trajectory, is the foundation of the HSI/ARI data sub-system. And without good position data, that sub-system is a house of cards.

When we began researching the ES evaluation problem several years ago, we anticipated access to army instrumented ranges. We soon discovered that such ranges were few, far between, expensive to operate, and mostly unavailable. As a result, we began a small study of low-cost portable alternatives to such facilities as those at the Combat Developments Experimentation Command and at Ft Hood. The study was particularly geared to supporting the upcoming Ft Carson test. The study, by Behavior Technology Consultants, Inc., looked at optical triangulation, optical ranging, unattended ground sensors and a number of radio ranging techniques. It recommended a radar ranging system which appears to be portable, relatively low cost and sufficiently accurate but which is still beyond our resources and which could not be put together in time for Carson anyway[5]. As a result, we are developing some labor intensive strategies for plotting tactical movements. These will involve intensive map and terrain training of data collectors, and systematic cross checking of results. In addition, the HSI methodology includes some techniques for screening bad position data and estimating missing points.

If we can succeed in adequately solving the dilemmas posed by Steps 1, 2, and 3 of an ES evaluation system we will have achieved a great

[5]O'Heeron, M.K., Howell, W.Y., Frazier, T.W., and Johnson, E. Field Measurement and Data Collection System for Engagement Simulation Field Exercises. Final Report. Behavior Technology Consultants, Inc. 1 October 1977.

FIG 5: EXAMPLES OF MEASURES OF EFFECTIVENESS

EXERCISE PHASE: PRE-ENGAGEMENT

MEASUREMENT CONCEPT: TACTICAL FORMATION

MOE: WHAT NUMBER OF SUB-DEPLOYMENTS (PARTITIONS) ARE ORDERED AND/OR INITIATED IN TEAM MOVEMENT TO CONTRACT?

WHAT SPECIFIC SUB-DEPLOYMENTS OF WHAT SPECIFIC FORCE MIXTURES WERE MADE WHERE, WHEN, AND MAINTAINED FOR HOW LONG? UNDER WHAT CONDITIONS?

EEAs: GROUP IDENTIFIERS, TRAJECTORY-TRIBUTARY NUMBER, FORCE MIXTURE, TEAM IDENTIFIER, HEX NUMBER, AND TIME.

FIG 6: EXAMPLE OF MEASURE OF EFFECTIVENESS

EXERCISE PHASE: ENGAGEMENT

MEASUREMENT CONCEPT: DETECTION OF ENEMY FORCES

MOE: WHAT TARGETS WERE DETECTED BY WHOM, WHEN, IN WHAT SUB-UNIT, HAVING WHAT FORCE MIX, IN WHAT SUB-DEPLOYED PARTITION, HEADED WHERE?

EEAs:

TARGET TYPE,
TARGET NUMBER,
DETECTOR TYPE,
DETECTOR NUMBER,
TIME,
LOCATIONS OF TARGET,
LOCATION OF DETECTOR.

deal. However, we still won't have, in my opinion, a criterion-referenced system. Being able to collect various kinds of process and product data, and being able to relate these data to each other is very critical. But their interpretation and usefulness for training diagnosis is incomplete without performance benchmarks or standards. The problem of standards has been avoided for two-sided combat training exercises because such exercises are situation-specific and involve a very complex and not well understood set of variables.

Accordingly, we have underway a basic research program to explore the standards problem. This effort stems, in part, from a model developed several years ago by Litton under contract to ARI. This model, the UNIT PERFORMANCE ASSESSMENT MODEL (UPAM) used the policy capture technique to generate indices of combat proficiency. The index values resulted from a linear combination of variables, mostly reflecting various casualty measures. Commanders' forecasts of these measures for an upcoming exercise were to provide benchmarks against which actual data were to be compared.

A more recent modeling effort is called COMBAT OPERATIONS TRAINING EFFECTIVENESS ANALYSIS (COTEAM). This effort picks up on the concept of situation-specific forecasting to provide performance benchmarks, but develops the concept further in some ways we think are quite significant.

COTEAM hopes to do several key things:

(1) Define methods for forecasting products, processes, and dynamics of ES exercises in a situation-specific context. The UPAM system required forecasting but left the problem up to the COs own devices.

(2) Partial out non-training effects such as terrain, mission type, force ratios and to do so by addressing two kinds of benchmarks

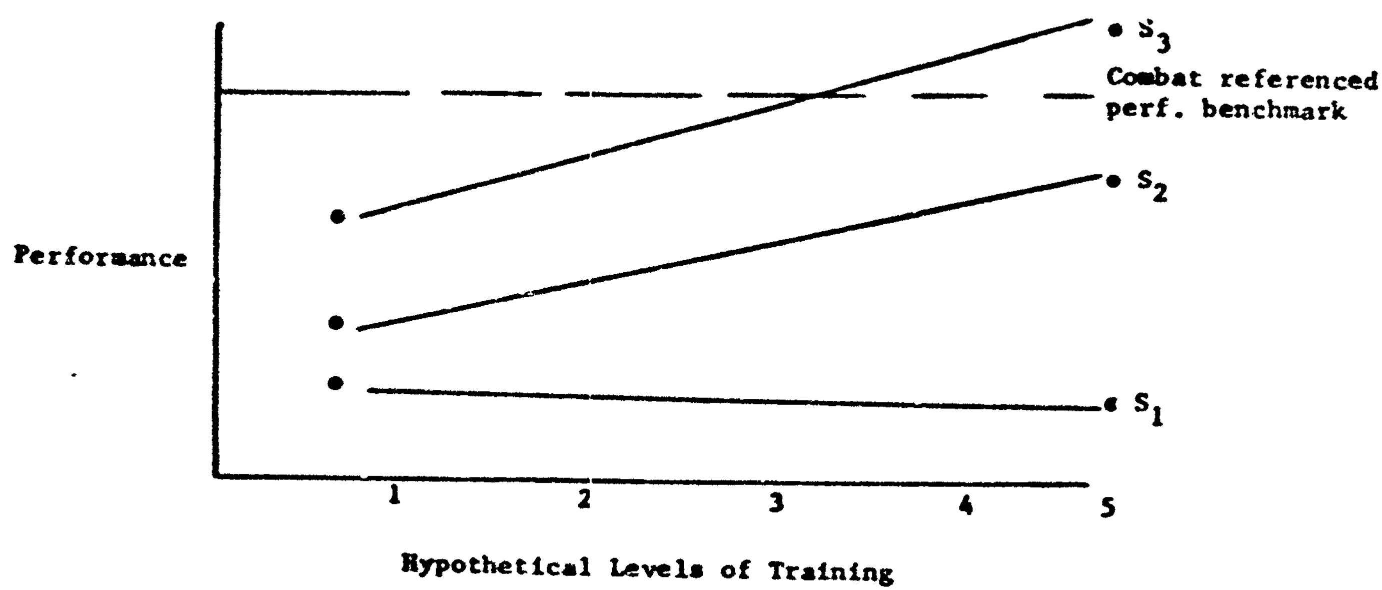
- o Training System Referenced
- o Combat Referenced.

The curve in Fig 7 suggests why you need two types of benchmarks.

Imagine that you could unambiguously define training system benchmarks (1-5) representing various points in a training cycle, e.g., 1 = entry level performance, 5 = final stage in unit training. Now imagine that you could define various sets of operational conditions (S_1, S_2, S_2) where such variables as weapon mix, terrain type, doctrine, force ratio define the sets. Further imagine that we could generate performance curves as functions of training level and operational condition set. What conclusions could you draw from Figure 7:

(1) The conditions of S_1 are such that training effects are completely overwhelmed. As a training manager you would avoid S_1 since it does not allow for a differentiation across levels of training.

FIG 7: EXAMPLE OF PERFORMANCE BENCHMARKS REFERENCED TO TRAINING SYSTEM AND COMBAT



(2) S_2 and S_3 on the other hand would both be potentially useful to the training manager.

(3) The individual who sits above the training manager, and balances training, force development, and doctrine would probably exclude both S_1 and S_2 as potential training conditions because neither set of conditions permits training to bring combat units up to an acceptable level of readiness.

But, how do we go about actually generating the benchmark curves? The answer to that question is the subject of current research. We have expectations of developmentally testing some techniques at Carson in January, e.g., the Delphi Technique as a way of systematically extracting predictions from experts. Other possible techniques would be combat board games and computer simulations. We collected some preliminary data on forecasting during a developmental test of rifle squad engagement simulation last April at Ft Ord without benefit of the Delphi method or board games. We wanted to get a feeling for the kind of data which might result.

Fig 8 describes the scenario and instructions which were given to subjects for a squad movement to contact. The subjects were NCOs acting as squad leaders.

Fig 9 shows the kinds of forecasting that was done by the NCO's and the data that resulted. Figs 10 and 11 show scenario, instruction, and resulting data for a hasty defense. Generally, our impression was that forecasting could be done with some reliability and that the task of forecasting for different assumed training levels was not an unsurmountable one. Our subjects did seem to be able to discriminate expected tactical performance across assumed training levels.

I've tried to provide a broad and very surface view of a complex research program. I really have not done justice to the scope of effort involved. Some indication of the size of the effort is its staffing. Approximately 16 ARI behavioral scientists with advanced degrees are partially or fully involved with the program. They are supported by the services of four private behavioral research companies. The very active and indispensable support of our TRADOC sponsors probably adds another five professional man years[6].

What the pay-off for this effort will be, I cannot predict. But its significance lies at least partially in its potential contribution to the Army's proposed National Training Center at Ft Irwin, California. Large sums of money are likely to be invested in the production of a very sophisticated instrumented range, capable of generating enormous quantities of high resolution data. If the Army's capacity to select, process, and interpret those data for training purposes does not match its capacity to supply the hardware and engineering involved in instrumenting a range, the Ft Irwin concept may not reach its full potential.

---

[6]Our sponsors are the Training System Manager for Tactical Engagement Simulation Systems, Ft Eustis, and The Directorate of Training Developments, Ft. Knox.

FIG 8: MOVEMENT TO CONTACT AGAINST AN OP

10 man attacking squad (testing squad)

4 man (standard) defense

Scenario: A ten-man squad is the point element of the platoon in a movement to contact. The squad will know that they can expect contact at any moment. They will have just crossed a danger area where they encountered sniper fire, without taking any losses. The squad is now approaching an enemy OP, consisting of four men with a machine gun in well concealed positions. Time $t = 0$ occurs as the squad clears the danger area.

Instructions

Your own opinions and estimates are being requested. This is NOT a test of your personality; the data will be used strictly for scientific purposes.

Assume that all members of your squad have only been through Basic Combat Training (BCT). Now on the next two pages, go down the first (BCT) column, and put your answer in each box for each question. If a more detailed answer is called for, use the reverse side of the paper.

Now assume that your squad has recently passed Infantry Level 2 ARTEP, shown in column two. Answer each of the questions again for this column.

Assuming three days of SCOPES training, answer the questions again in the third column.

Finally, assume that all members of your squad are combat experienced Rangers, and answer all questions in the boxes for the fourth column.

FIG 9: FORECASTS OF INFANTRY SQUAD MOVEMENT TO CONTACT

Level of Training of Tested Squad

QUESTIONS ASKED	Only B.C.T.	Passed Level 2 Infantry ARTEP	Passed Level 2 ARTEP, and 3 or more days of SCOPES Training	Combat Experienced Rangers
1. Maximum distance (meters) between fire teams.	11.3# 7.4*	19.6 8.7	21.7 12.5	34.3 18.4
2. Minimum distance (meters) between fire teams.	7.6 7.8	17.7 13.7	20.5 15	30.8 14.8
3. How many meters will the point man be in front of the squad?	11.8 10	21.9 14.8	26 14	40 6 23
4. What will be the distance (m) between the right and left flanks?	18 14	31 20	34 20.4	42.5 27.8
5. Will the OP be detected prior to anyone crossing Phase Line C?	no - 7 yes - 1	no - 4 maybe - 4	no - 2 maybe - 6	no - 2 yes - 6
6. What will be the approximate range at initial detection by the OP? (m)	88 19	71 40.4	80 54.5	62.6 53.2
7. Will the OP be detected prior to anyone opening fire?	no - 7 yes - 1	no - 2 yes - 2 maybe - 4	no - 1 yes - 4 maybe - 3	no - 1 yes - 7
8. How many minutes do you think will elapse before the squad will be detect by the OP?	5.3 5.8	7.1 8.2	7.4 9.8	14.6 13.3
9. How many OP casualties do you think there will be?	2 1.6	1.7 .9	2.5 .8	3 1.5

#--top value is the **mean** of 8 respondents; * is the standard deviation from that mean.

FIG 9: FORECASTS OF INFANTRY SQUAD MOVEMENT TO CONTACT (CONT)

QUESTIONS ASKED	Level of Training of Tested Squad: Only B.C.T	Passed Level 2 Infantry ARTEP more	Passed Level 2 ARTEP, and 3 or more days of SCOPES Training	Combat Experienced Rangers
10. What per cent chance do you think the squad has of taking the OP?	17.8 21.2	45 23	68 11.1	88.4 13.2
11. How long do you think the exercise will last if the squad takes the OP? (min)	37.5 24	32 15.6	21.4 13.2	27.6 21.4
12. If we divide the duration of this exercise into quarters, how many casualties do you think the squad will sustain during the first quarter?	2.9 1.4	2.2 .5	1.5 .5	1.3 1.1
13. How many casualties during the second quarter?	1.8 .5	.5 1	1.4 1	1.4 1.5
14. During the third quarter?	1.9 1.1	1.5 1.1	1.1 1.2	.9 1.3
15. During the last quarter?	2.3 .8	1.9 1.5	1.8 1.5	1.6 1.9

FIG 10: HASTY DEFENSE

7 man attacking squad (controlled aggressor)

10 man defending squad (tested squad)

Scenario: A ten-man squad established a hasty defense as part of a larger platoon defensive perimeter. They will have approximately 15 minutes from the delivery of the frag order to establish the hasty defense. At that time an enemy counterattack, consisting of 7 men with a machinegun will begin their approach toward the defensive positions. The counterattack movement will begin at a position approximately 100 meters from the defense. Time t = 0 occurs with the delivery of the frag order.

Instructions

Your own opinions and estimates are being requested. This is NOT a test of your personality; the data will be used strictly for scientific purposes.

Assume that all members of your squad have only been through Basic Combat Training (BCT). Now on the next two pages, go down the first (BCT) column, and put your answer in each box for each question. If a more detailed answer is called for, use the reverse side of the paper.

Now assume that your squad has recently passed Infantry Level 2 ARTEP, shown in column two. Answer each of the questions again for this column.

Assuming three days of SCOPES training, answer the questions again in the third column.

Finally, assume that all members of your squad are combat experienced Rangers, and answer all questions in the boxes for the fourth column.

FIG 11: FORECASTS OF SQUAD HASTY DEFENSE

Level of Training of Tested Squad

QUESTIONS ASKED	Only B.C.T.	Passed Level 2 Infantry ARTEP	Passed Level 2 ARTEP, and 3 or more days of SCOPES Training	Combat Experienced Rangers
16. Describe, on the enclosed sheet of paper, what you think the defending squad's reaction to attack will be.				
17. The overall tactical performance of the defending squad will be, on a scale from 1 (excellent) to 7 (extremely poor):	6.# 1.*	4.3 .1	3.4 1.4	2.5 2.3
18. The overall control of the squad during this exercise will be from 1 (excellent) to 7 (extremely poor):	5.4 1.5	4.2 .1	3.7 1.4	2.9 2.6
19. If detection is made before contact was initiated, what will be the range of detection? (m)	85.5 62.2	119 62.8	136 63.7	158 86.2
20. What will be the range of the OPPFORS when contact is initiated? (m)	76.8 66.5	80.2 75.7	82.8 70.7	88.3 91.4
21. Which side do you think will open fire first?	Defense: 6 Attacker: 2 ? = 1	Defense: 6 Attacker: 2 ? = 1	Defense: 7 Attacker: 1 ? = 1	Defense: 8 Attacker: 1
22. At what range do you think the OPPFORS will be detected? (m)	46.4 79	80. 72.5	81 76	110 113
23. What percent chance do you think the squad has to hold its defensive position?	28 32	59 23	56 20	84 19

#--top value is the mean of @ respondents; * is the standard deviation from that mean.

FIG 11: FORECASTS OF SQUAD HASTY DEFENSE (CONT)

Level of Training of Tested Squad

QUESTIONS ASKED	Only B.C.T.	Passed Level 2 Infantry ARTEP	Passed Level 2 ARTEP, and 3 or more days of SCOPES Training	Combat Experienced Rangers
24. How long do you think the squad will hold, before (if) they are overrun? (min)	12.1 13.7	26 19.9	32.6 28.3	34.8 24.5
25. How long do you think this exercise will last if the squad successfully defends their position? (min)				
26. If we divide the duration of this exercise into quarters, how many casualties do you think the squad will sustain during the first quarter?	2	2	1	1
27. How many during the second quarter?	2.5	2	1	1
28. During the third quarter?	2.5	2	1	1
29. During the last quarter?	2	2	1	1

RELIABILITY IN MEASURING UNIT PERFORMANCE

John A. Boldovici
William C. Osborn
and
James H. Harris

Human Resources Research Organization
Fort Knox, Kentucky

Paper for:

Military Testing Association
San Antonio, Texas

October 1977

RELIABILITY IN MEASURING UNIT PERFORMANCE

A central problem in all evaluations, and especially in evaluations of combat units, is how to incorporate characteristics of good measurement in the evaluations. Characteristics of good measurement include comprehensiveness, cost-effectiveness, validity, and reliability. Our concern in this paper is with reliability; for without reliability, comprehensiveness is of little value, and cost-effectiveness and validity cannot be achieved.

Reliability refers to the extent to which:

1. Two or more independent observers produce similar results, and

2. Measures of an event taken at one time are identical to measures of the same event taken at another time.

The performance of combat units, at least in the Army, is increasingly being evaluated in the context of large-scale, free play simulated combat exercises. The ARTEP (Army Training and Evaluation Program) is an example. The results of performance evaluations in simulated combat are used by policy makers in decisions about training needs and combat readiness. Given the importance of decisions about training needs and combat readiness, and given the dependence of these decisions on unit performance evaluations, a question naturally arises as to how to maximize reliability in measuring unit performance.

Purpose

The purpose of this paper is to present hypotheses about variables that affect the reliability of unit performance measurement, and to outline research for testing the hypotheses.

Sources of Measurement Reliability

Measurement can be viewed as consisting of three phases:

1. Observer Preparation.

2. Observation.

3. Recording and Reporting.

Variables that affect measurement reliability are at work within each of the three phases of measurement -- variables that affect the extent to which two or more observers produce similar measurement results, and the extent to which measures taken at a given time are representative of measures taken at another. Hypotheses about the variables in each of the three measurement phases follow.

Observer Preparation

Reliability of measurement will increase with the consistency or uniformity of understanding among observers about the rules of observation and recording. Ideally, observers should be standardized, and measures should be taken to assess the degree to which they have been standardized. Measurement reliability may be increased by manipulating the following variables in the observer preparation phase:

1. Specificity of instructions. Reliability is likely to be greater when the instructions to observers are highly specific than when instructions are general and loosely stated.

2. Timing of instructions. Instructions to observers should not be given so far in advance of observation as to permit forgetting, or so late as to preclude learning.

3. Practice in observing and recording. Measurement reliability will be greater when observers have practice measuring and recording the events of interest than when they have not. The practice variable interacts with timing of instructions, in that instructions to observers should be given far enough in advance of observation to allow time for practice.

4. Testing observers. Measurement reliability can be indirectly increased by the use of tests to make sure that observers are capable of performing whatever measurement operations will be required of them.

Observation

Even with very careful observer preparation and totally standardized observers, measurement reliability will be affected by variables at work during the observation (measurement) process.

Properties of the events or things to be measured can affect measurement reliability. Measurement of unidimensional events will, for example, be more reliable than measurement of multidimensional events (all other things being equal). This is related to perceptual "clutter," or limits on observers' information-processing abilities. Within rather broad limits, observers who are asked to make large numbers of simultaneous observations and measures will produce less reliable results than will observers making smaller numbers of observations.

Another property of the events or things to be measured that affects measurement reliability is stability (or its opposite, transience). The results of measuring the diameter of a wooden ball will, for example, be more reliable than will the results of measuring a mercury "ball" -- once again, all other things being equal.

Other properties of events to be measured that will influence reliability are time-sharing, noise, and "observability"; that is, measurement reliability may be expected to decrease with the extent to which the observed event is:

1. Time-shared with other events.

2. Embedded in noise.

3. Not directly observable.

Strategies, rules, and procedures for measurement also affect reliability. Observers may be expected to perform more reliably, for example, to the extent that they are:

1. Required to make comparative rather than absolute judgments.

2. Given a well defined standard stimulus.

3. Alerted as to what to observe (anticipate likely errors).

4. Given the opportunity to observe an event more than once.

5. Given scoring aids or templates.

6. Required to measure only, and not process measurement results.

Recording and Reporting

Even with adequate observer preparation and careful control of the measurement process, measurement reliability will be affected by variables operating during the recording and reporting of measurement results. These variables include:

1. Timing. Measurement reliability will increase with decreased time between observation of the event of interest and recording of results.

2. Design of recording forms. Well designed data recording forms minimize the amount of judgment and decision-making required for their use, and thereby increase the reliability of recorded results. Simplicity in data-recording forms, for example, minimize data-recording time, and therefore allows more time for observation.

Unit performance measurement probably is unreliable because of the influence of all of the variables mentioned above. These variables serve to decrease the reliability of operations as simple and straightforward as measuring length with a ruler. The considerable complexity of free play simulated combat guarantees that measurement reliability problems will be great.

In the observer preparation phase, for example, observers may not be standardized for any number of reasons. Instructions for measurement may be too general, and may not be given at the right time Observers may not have enough practice to permit performing their measurement duties in accordance with the intent of the test designers. And practical constraints (e.g., time, money) may preclude ascertaining whether observers are capable of performing their measurement duties before "turning them loose."

In the observation phase, observers may be required to make simultaneous judgments along more dimensions than their sensory apparatus can comfortably handle. The measurement instruments may permit too much subjectivity and expertising. Strategies for measurement may be inappropriate (single rather than multiple observations, for example). And the nature of the required judgments and decisions may invite unreliability.

In the recording and reporting phase, unreliability may be promoted by the length of time between observation and recording of results, and by formats for recording results.

The possible influences of the variables discussed above demand that research be undertaken on methods for improving the reliability of unit performance measurement, for measurement without reliability will lead to wrong decisions about training needs and about readiness.

Photography and Measurement Reliability

The conduct of measurement reliability studies requires that whatever is to be observed and measured (simulated combat, for example) must:

1. "Sit still" long enough to permit observers to make the required measures.

2. Be presented uniformly or varied systematically for various groups of observers.

These two requirements, and the high cost of field studies using simulated combat, make the conduct of field studies of measurement reliability impractical. The requirements for "sitting still," for uniform or systematically varied presentation, and for low cost can be met by the use of photography.

Motion pictures of simulated combat can be made, using real combat vehicles or models. Models seem preferable for two reasons. The first is low cost. The second is that research on reliability of measuring unit performance does not require perfect fidelity or realism in the events to be observed and measured. As noted earlier, the main requirement is for a set of events that can be presented uniformly to various observers, or varied in accordance with requirements of the experimental design.

Subtle errors in tactics and operations can be deliberately incorporated into motion pictures, for the purpose of producing variability in observers' response to events presented in the film. And by editing videotape versions of the film, the amount of information available to various groups of observers can by systematically varied.

Studies of reliability in unit performance measurement should take the following general form: A set of events is selected for observation and measurement (e.g., a part of the ARTEP). Several groups of subjects view the events, observing, measuring, and evaluating according to instructions and experimental conditions. Systematic variations are introduced in variables in any or all of the three phases of measurement. As implied earlier, variations could be introduced in the kinds of instructions given to observers, the specificity of the instructions,

amount of practice given to observers, kinds of instruments and measurement strategies, and so forth. In all cases the dependent variable is an index of inter-observer reliability; e.g., a simple "percent-agreement" score to indicate the extent to which observers produce similar results measuring the same things. Variables that affect reliability are identified, and can be incorporated into "how-to" literature for reliable unit performance measurement.

The conduct of research along the lines suggested above seems warranted, because the results would lead immediately to action recommendations for improving measurement reliability, and could be incorporated directly into any program for measuring unit performance.

PERSONNEL TURBULENCE AND TIME UTILIZATION IN AN INFANTRY DIVISION*

Dr Hilton M. Bialek
Diana Zapf
Wendee McGuire

Introduction

In its attempts to comply with recent DOD policy "that learning objectives which can be accomplished more economically in the operational unit, and without unacceptable degradation of unit readiness, should be provided as OJT rather than as individual training", the Army has instituted a number of R&D efforts designed to decentralize training. A number of these efforts, and one I have been involved in for the past two years, utilizes the squad leader as a primary instructor. The idea, in addition to decentralization, is to enhance the leadership role of the squad leader by making him primarily responsible for the individual skill proficiency of the men under his command.

For an instructional system like this to work, some sort of personnel stability would seem necessary. A squad leader needs sufficient time to learn the strengths and weaknesses of his men, time to create a group identity and cohesion, and time, of course, to provide instruction. How stable then, are TO/E companies and squads? That is one question we attempted to answer. The second question had to do with the utilization of time: "How much time does a squad leader typically have to actually devote to training?" These two questions guided the design and conduct of the study I will now describe to you.

Approach

To investigate these questions, two main sources of information were used.

- The manning reports submitted monthly by each company to the battalion headquarters.
- A large sample of 15-minute-by-15-minute first-hand observational records of the daily activities of individual squad members.

Manning reports described the flow of personnel in and out of the four companies in the sample, as well as the duty positions and MOSs of each man in the company. The second data source — observations — pro-

*Paper presented at Military Testing Association Conference, San Antonio, October 1977.

vided information concerning what men in a sample of squads selected from the four companies were doing on a quarter hour by quarter hour basis and how long they were doing it.

The operational unit in this instance was a CONUS infantry division. Observations and sub-unit sampling focused on 11B and 11C MOSs because these MOSs had been selected as the initial content of the individual skill training system under development.

The Sample

The sample selected reflected our interest in studying turbulence on both the individual and unit level, for MOSs 11B and 11C. The sample of companies and squads within those companies was chosen to represent:

SLIDE 1 HERE

Results

I will report first the analysis of the data available from the manning reports and the accompanying information obtained by checking battalion records. Later I will focus on the data obtained from the daily observation phase of the study. Movements in and out of companies and squads is shown in Slide 2.

SLIDE 2 HERE

Slide 2 shows that after 4 months:

24% were no longer in the company.

24% had moved to another squad within the company.

In terms of stability, at the squad level, we see that:

52% were in the same squad.

36% were in the same duty position in the same squad.

16% were in different duty positions in the same squad.

As backup data for this manning report information, the observers we employed to measure time utilization were required to record the actual names of squad members each time they spent the day with the squad. Averaging the results from observing 10 squads gave the results shown in Figure 3.

SLIDE 3 HERE

The results show that 31% of the original squad members left the squad over a *two* month period. Assuming a linear relationship, there is about a 15% turnover per month. Comparing this to Figure 2 — 48% left the squad over a four month period giving a monthly rate of 12% — shows the estimate to be quite close. Also shown are the number of "movements" — the arrival or departure of a member — experienced for the composite (average of the ten squads observed) squad. The total number, 8.67 is perhaps a more sensitive turbulence indicator than simple turnover (the proportion of positions initially observed that are held by someone else at the end of the observation period) because it includes individuals who arrive and depart between the initial and final observation points.

We will reserve comment on these findings until after we have presented our time utilization results.

The second part of the study — measuring daily turbulence and time utilization — involved a direct observation technique. Four volunteers from an engineering battalion were trained as observers. Each observer was assigned to one of the four companies comprising the sample. Each day the observer would meet with the company at morning formation and spend the remainder of the duty day with a designated squad[1], following them wherever they went. The observer carried a clip board with a copy of the observation data sheet attached, and every 15 minutes recorded the activity of each member of the squad on the basis of two decisions: (1) which of six major activity areas is the soldier involved in? and (2) within that activity area, which of four modes is he in? The six activity areas are described in Table 1, along with examples.

SLIDE 4 HERE

Within each of these activity areas, the observed person was categorized into one of four modes: receiving information, performing tasks, waiting to receive instruction, or enroute to or from activity.

During the training of observers, reliability checks were conducted by having two observers observing the same squad for a full day. For over 90% of those time-units observed and recorded, both observers recorded the same categories. The differences that did occur were mostly in the "mode" dimension, whereas the major activity areas seemed clearly discriminable.

[1]On those occasions when the unit was engaged in night training exercises, the observer would spend the night observing the squad.

Using this data collection procedure, the activities of a total of 166 observer days or nearly 40,000 15-minute time units were recorded. I will now talk about how these 40,000 time units were distributed.

SLIDE 5 HERE

The results are organized to show how, for a typical or average training day, non-training day, and overall average day (the two types of days combined) time and people are utilized. Slide 5 shows the total number of time units observed, broken down by number and percentage. Each time unit represents one man for a period of 15 minutes. For example 2901 time units were recorded as "unit training" during the training days. This number is 14% of the 20,626 time units observed during training days. Note that a category labeled "absence" (turbulence) is included in the total number of time units. This is the number of time units *lost* because individuals who were officially available for duty were not in fact present. These absences could range from one time unit during the day to the entire day. The numbers shown in Table 2 indicate that 18% of the time units available during training days (15% during non-training days and 16% overall) were unused because of absences.

Perhaps a more direct way of portraying the results is to show what the typical infantryman spends his time doing in a typical duty day, and how much time he spend doing it.

SLIDE 6 HERE

The next slide shows this for a training day; I'll not load you down with too much by showing the non-training and combined day but obviously they show less training time.

Turning to the question of what soldiers do when they are absent from the squad (on the average of one hour 23 minutes per day) a breakdown of their activities appears in the next slide.

SLIDE 7 HERE

Results are, again, shown for a training day, a non-training day, and both days combined. The major turbulence causing activities seem to be work details (27% of the absences during a training day are a results of this) and military schooling (another 21%). The remainder of the time is a result of the other activities listed on the slide. Absences from the squad occur as frequently during actual training time (unit and individual) as during other activities.

One final analysis shows that of the two hours seven minutes designated as training (unit and individual) spent on a "training" day, the average soldier spends 57 minutes of that time actually engaged in hands-on performance behavior. He spends another 25 minutes per day receiving instruction. The last slide

SLIDE 8 HERE

shows how the remainder of his actual training time is distributed.

Comments

First a methodological comment. I think this direct observation technique for obtaining data is highly useful and can be applied any time one wishes to find out whether or not organizational policy changes do in fact change patterns of action and time utilization. It is not difficult to train observers and the system is hardly affected by their continuous and extended presence.

Turning now to the results themselves --

Although awareness of turbulence and efficient use of time is recognized and widespread, it seems that this close look at a 3-4 month period in the life of an operational unit is less than comforting. During that period, overall, less than 25% of the time is actually devoted to training and of that time about one-third is lost to delays, movement, and other minor factors.

Certainly the amount of movement of personnel both within and into and out of an operational unit appears excessive and the obvious question is, "Is it all necessary?" Are there personnel management policies in the Army which contribute to this movement? Are commanders usually short-handed and therefore needing to shuttle people around to

fill in gaps on a temporary basis? Are administrative and support requirements given greater priority than training/operational requirements? Is the inefficient use of time reported above a consequence of policy decisions, a breakdown in line of command, or due to yet other causes? In other words, is the picture conveyed in this study inevitable or can certain elements be identified which, if modified, would change the pattern of movement and time usage described above.

If may well be, however, that the pattern revealed in this particular kind of operational unit — an infantry division — is not found in other kinds of Army units, operational or otherwise. There may be something special about combat arms (or infantry specifically) that exacerbates the problem. For example, combat oriented operational units are in the unique position of never (except in actual combat) having to perform the jobs they are trained for. Thus, they are always in "training" or in a state of preparation. The distinction between such a unit's operation and its preparation or training for that operation is, at best, fuzzy. If may well be that this unusual circumstance leads to greater turbulence and inefficient time usage as compared to other operational units (a transportation unit, an administrative center — many combat support activities). Another example of a factor that might contribute to turbulence is the *actual* priority training is given, vis a vis the other demands placed on an operational unit; housekeeping, maintenance, unit missions, officer career requirements, etc. The point of this discussion is that a case can be made for investigating the causes and conditions that relate to turbulence and time utilization because (a) it is highly likely that these two phenomena are related to organizational effectiveness and efficiency, and (b) the factors underlying them, once identified and isolated, can probably be greatly modified so as to improve effectiveness and efficiency. It would appear therefore that the need for further direct investigation into the causes and amelioration of these phenomena is warranted while attempts to design training systems which can handle them continues.

SLIDE 1

1. Brigades: Two companies each from the division's two maneuver brigades.

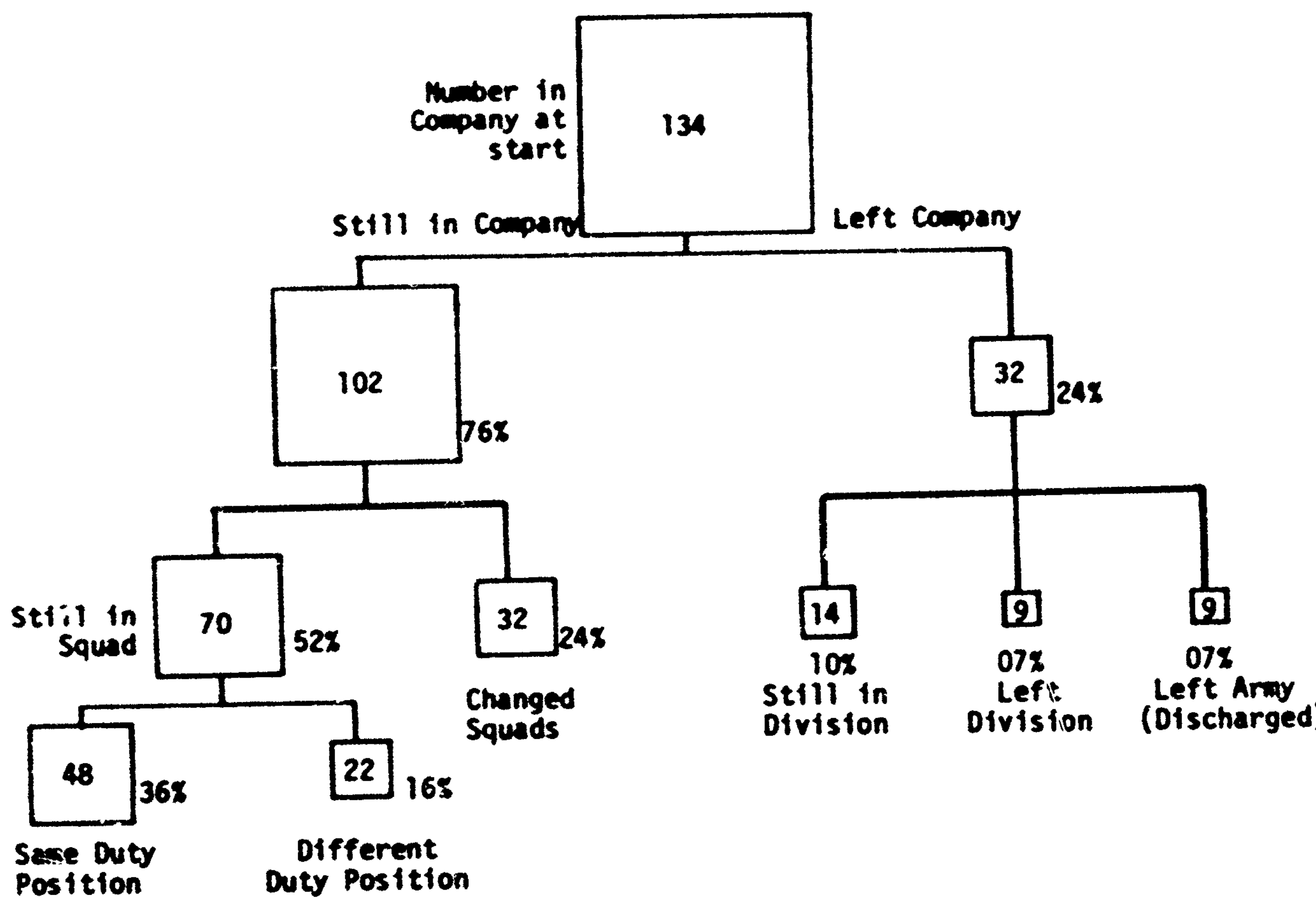
2. Battalions: Four of the six maneuver battalions were represented in the sample.

3. Kinds of companies that contain most of the 11Bs and 11Cs: Rifle and combat support companies were studied.

4. The existing ratio of 3 rifle companies to 1 combat support company: 3 of the companies in the sample were rifle companies, the fourth was a combat support company.

5. Kinds of platoons: 10 different platoons within the 4 companies were studied, one squad from each platoon.

DISPOSITION OF MEN IN A COMPOSITE RIFLE COMPANY
AFTER FOUR MONTHS



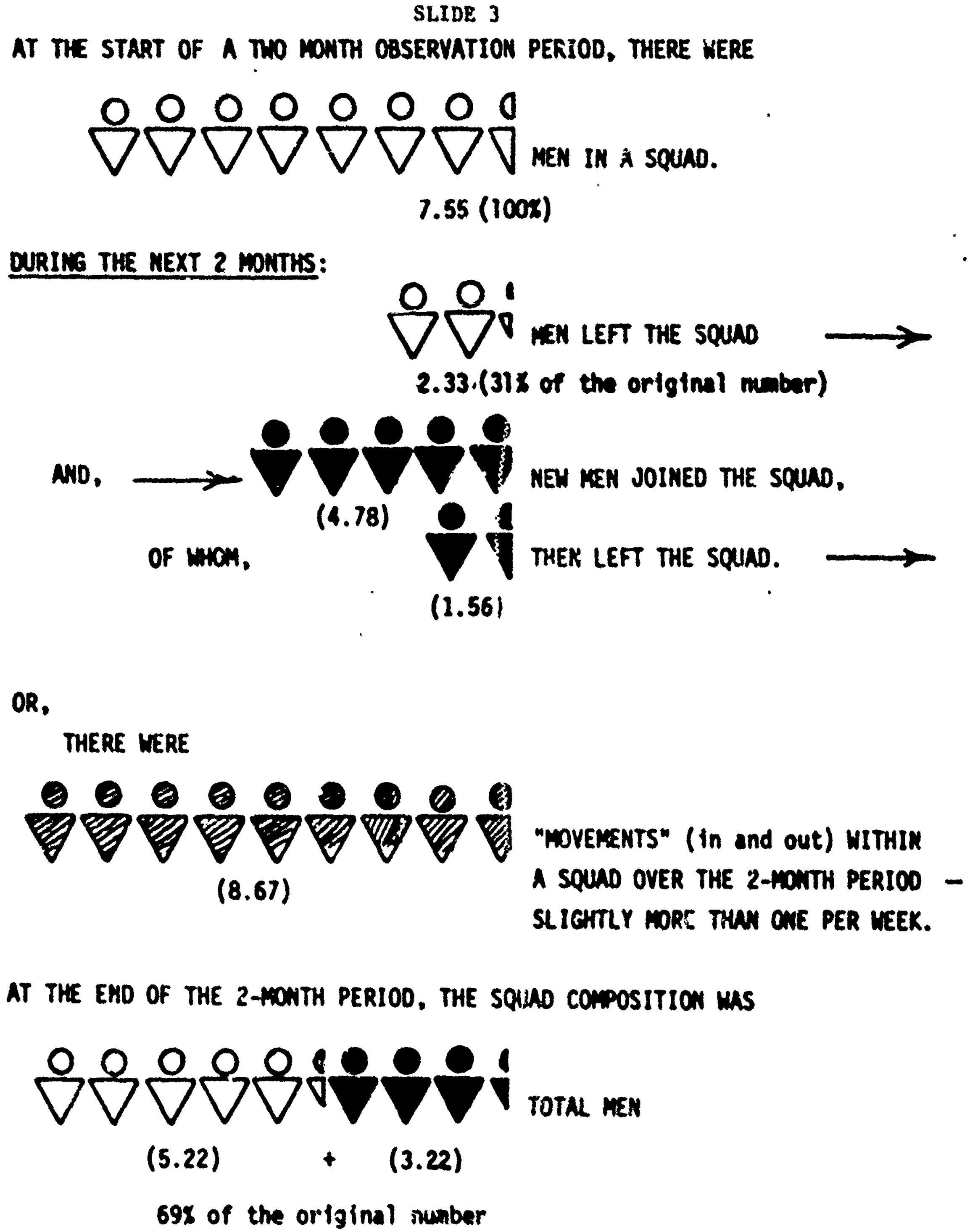


Figure 2. Composite Squad Level Turbulence.

SLIDE 4

TABLE 1. MAJOR ACTIVITY OBSERVATION CATEGORIES

ACTIVITY CATEGORY	EXAMPLES
UNIT TRAINING (U) Focuses on training individuals to perform as members of a team or unit.	ARTEP Field Exercise: squad ambush. Indoor class on assembly area procedures. Field Exercise: company defense.
INDIVIDUAL TRAINING (MOS Skills) (I) Focuses on the skills (tasks which the individual needs to do his job.)	Weapons qualification. Indoor class on camouflage techniques. Outdoor class on mine detector training. EIB training. Mortar Crew drill. Class on first aid.
INDIVIDUAL TRAINING (PT) (Ipt) Physical readiness training.	PT Unit team athletics.
TEACHING ACTIVITIES (T) Teaching or assisting in teaching for unit or individual training.	Teaching a class on land navigation. Demonstrating how to set up a minefield.
SUPPORT/GARRISON (S) Activities which support training; garrison duties.	Weapons issue and turn-in. Maintenance of weapons, equipment, vehicles. Maintenance of billets/buildings. Work details. Parades Garrison guard mount. CQ
PERSONAL CARE (P) Authorized activities only.	Breaks Taking showers. Changing clothes.

SLIDE 5

TABLE 2. DISTRIBUTION OF TOTAL TIME UNITS BY MAJOR ACTIVITY AREAS.

ACTIVITY AREAS	TRAINING DAYS		NON-TNG DAYS		ALL DAYS	
	NUMBER OF TIME UNITS	% OF TNG DAY TOTAL	NUMBER OF TIME UNITS	% OF NON-TNG DAY TOTAL	NUMBER OF TIME UNITS	% OF ALL DAYS TOTAL
UNIT TNG	2901	14%	850	04%	3751	09%
INDIV TNG	3078	15%	782	04%	3860	10%
INDIV TNG (PT)	1722	08%	1982	10%	3704	10%
SUPPORT/GARRISON	6297	31%	10621	56%	16918	43%
PERSONAL CARE	2832	14%	1971	10%	4803	12%
TEACHING ACTIV	139	01%	141	01%	280	01%
ABSENCES (TURB)	3657	18%	2778	15%	6435	16%
TOTALS	20,626		19,125		39,751	

NOTES

1. Average number of men per Squad: Training = 8.03
Non-Training = 8.46
All Days = 8.25

2. Time units are recorded from the official start of the duty day to the official close. Lunch time is NOT included as a recorded time unit.

MAJOR ACTIVITY	TIME SPENT	% OF TOTAL DUTY DAY
UNIT TRAINING	1 hr 02 minutes	14%
INDIVIDUAL TRAINING	1 hr 05 minutes	15%
PHYSICAL TRAINING (PT)	37 minutes	08%
TRAINING OTHERS	03 minutes	(01%)
PERSONAL CARE	1 hr	14%
SUPPORT/GARRISON ACTIVITIES	2 hrs 14 minutes	31%
ABSENT (TURBULENCE)	1 hr 18 minutes	18%
	TOTAL DAY = 7 hrs 19 minutes	

Figure 3. Distribution of Time Devoted to Major Activities During an Average TRAINING Duty Day.

SLIDE 7

TABLE 3. BREAKDOWN OF ACTIVITIES ENGAGED IN WHILE ABSENT FROM DUTY

	% OF TOTAL TIME ABSENT		
ACTIVITY	TNG DAY	NON-TNG DAY	COMBINED DAY
MEDICAL	10%	03%	07%
PERSONAL	04%	01%	03%
MILITARY EDUCATION	21%	28%	25%
PERSONAL EDUCATION	08%	04%	07%
DETAILS/CQ	27%	25%	26%
DISCIPLINARY	0	11%	06%
LEAVE	08%	11%	10%
CLEARING	10%	01%	06%
COMP TIME	07%	05%	07%
OTHER	03%	11%	03%
TOTAL TIME ABSENT	1 hr 18 min	1 hr 02 min	1 hr 13 min

SLIDE 8

TABLE 4. ANALYSIS OF BEHAVIORAL MODE DURING UNIT & INDIVIDUAL TRAINING ACTIVITIES FOR A TRAINING DUTY DAY

MODE	TIME	% OF TOTAL TNG TIME
RECEIVING INSTRUCTION	25 min	20
PERFORMING TASKS	57 min	45
WAITING TO RECEIVE INSTRUCTION	10 min	08
ENROUTE	23 min	18
OTHER	11 min	09
TOTAL	2 hrs 07 min	100%

SLIDE X

Of the 24% (32 men) who left the company:

9 — DISCHARGED	23 — TO OTHER UNITS
2 ETS	3 OVERSEAS
2 CHAPTER 13	2 OTHER DIVISION
2 CHAPTER 15	7 OTHER BATTALION
2 AWOL-DFR	7 OTHER COMPANY, SAME BATTALION
1 MEDICAL	1 SPECIAL SCHOOL
	3 RE-ASSIGNABLE OVERSTRENGTH

PROBLEMS IN MEASURING TEAM EFFECTIVENESS

Albert L. Kubala
Human Resources Research Organization,[1] Fort Hood, Texas 76544

Background

Borrowing heavily on characteristics of teams described by Glaser, Klaus, and Egerman,[2] as well as Hall and Rizzo,[3] Wagner, Hibbits, Rosenblatt, and Schulz[4] define team training as:

SLIDE 1

> The training of two or more individuals who are associated together in work or activity. The team is relatively rigid in structure and communication pattern. It is goal- or mission-oriented with the task of each team member well-defined. The functioning of the team depends upon the coordinated participation of all or several individuals. The focus of team training and feedback is on team skills (e.g., coordination), activities and products.

It can be seen from the implied definition of a team, that a team could be composed of anything from a two-man crew to a unit of almost any size. However, most of the literature dealing with teams has considered relatively small units such as those associated with one piece of equipment, such as a tank or aircraft, or at most, a platoon with a single objective or mission. Wagner, et al. further point out that, while the military

---

[1]This work was performed under Contract DAHC19-75-C-0025 to the US Army Research Institute for the Behavioral and Social Sciences (ARI). Dr. Charles O. Nystrom was the Contract Monitor.

[2]R. Glaser, D. J. Klaus, and K. Egerman. *Increasing team proficiency through training: 2. The acquisition and extinction of a team response*, Technical Report AIR B64-5/62, American Institutes for Research, May 1962.

[3]E. R. Hall and W. A. Rizzo. *An assessment of US Navy tactical team training: Focus on the trained man*, TAEG Report No. 18, Training Analysis and Evaluation Group, March 1975.

[4]H. Wagner, N. Hibbits, R. D. Rosenblatt, and R. Schulz. *Team training and evaluation strategies: State-of-the-art*, Technical Report 77-1, Human Resources Research Organization, Alexandria, Virginia, February 1977.

services conduct up to 90% of their training in the operational commands, most training research has been focused on individual training in institutional settings. For example, in FY 1974, the Army Research Institute for the Behavioral and Social Sciences (ARI) initiated the largest program of unit training and evaluation research in history. Yet, only 11% of the human resources budget was spent in this area. Judging from the literature, the resources devoted to this area by the other military services has been roughly comparable. This lack of emphasis seems strange in view of the fact that most fighting has been, and will continue to be done by teams. It now seems critical that we determine how well our teams do function, for as MG Gorman has stated, we must:

<u>SLIDE 2</u>

> ...train the Army to win on the first battlefield of the next war against an enemy that outnumbers us, against an enemy whose weapons will be as good as or nearly as good as those we possess....[5]

In other words, we can ill afford any but the most effective fighting teams. And, to ensure maximum effectiveness, Measures of Effectiveness (MOE) must be derived so that commanders can evaluate their own teams, discover deficiencies, and take corrective measures.

Our HumRRO contingent at Fort Hood became involved in this area when we were asked to determine what set of MOE were currently being employed to evaluate tank crews, and to determine what additional research was needed to ensure a comprehensive evaluation capability. We soon found that for all practical purposes, the only MOE in current use are scores on Table VIII, otherwise known as the Tank Crew Qualification Course.[6] For those of you unfamiliar with Table VIII, it should suffice for the moment to know that it is a live-fire gunnery exercise, where crews are scored on both hit accuracy and times to engage targets. Looking further at this MOE, we were surprised to find that the reliability of Table VIII scores has apparently never been determined, and that many question its validity as a predictor of combat effectiveness. We wondered why no other MOE were in use, and why one which was somewhat suspect was in general use. We wondered what the problem(s) was(were). Therefore, we decided that the

---

[5] W. E. DuPuy and P. F. Gorman. "TRADOC mission and resources briefing," transcript from TV tape, US Army Training and Doctrine Command, Fort Monroe, Virginia.

[6] J. A. Larson, W. K. Earl, and V. A. Hensor. *Assessment of US tank crew training*, TCATA Test Report No. FM 331, Final Report (23 March 75 - 13 March 76), HQ, TRADOC Combined Arms Test Activity, Fort Hood, Texas, July 1976.

next step should be a study of the problems associated with the development and use of team evaluations, which is the subject of this paper.

I doubt that anything I say will really be new to any of you. My purpose in presenting this paper is simply to re-focus your collective attention on these problems. I feel that the areas of team training and evaluation, especially evaluation, have been much neglected. Hopefully, this presentation will generate some interest in and lead some of you toward, solutions for some of the problems I will discuss. We have painstakingly developed procedures for building training programs and evaluating individuals. We have out inter-service procedures for instructional systems development,[7] and are now, in the Army, developing individual Skill Qualification Tests (SQTs). These tests will be designed to test actual job performance as well as knowledge, and successful performance will be a prerequisite for both retention and/or promotion. However, we have no similar procedures for either curriculum development or evaluation of teams, and they are sorely needed.

Problems

The particular problems which I have chosen for further elaboration are shown in the next slide.

SLIDE 3

- Defining Effectiveness
- Defining Team Effectiveness
- Problems With Numbers
- Reliability
- Evaluation Strategies
- Resources

Defining effectiveness. Historically, MOE were derived to ensure the quality of newly developed hardware. For one of our simplest weapons--the rifle--accuracy was the original MOE. Somewhat later, rate of fire was added as an MOE. Still later, it was realized that a highly accurate rapid fire weapon was of little value unless it were completely functional. Therefore, the concept of "availability" came into being as an MOE, and was measured by such things as Mean Time Between Failure (MTBF) and Mean Time to Repair (MTTR). However, the primary reason for the proliferation of MOE was the recognition that effectiveness was mission-dependent. For example, the weapon characteristics desirable for a sniper rifle are quite different from those required for a weapon designed primarily for suppression. In selecting a rifle, a sniper would be primarly interested in

---

[7]TRADOC PAM 3?0-30. *Interservice procedures for instructional systems development*, US Army Training and Doctrine Command, Fort Monroe, Virginia, 1 August 1975.

accuracy and range, but would not be too concerned about rate of fire. On the other hand, the soldier with the suppression mission would be very concerned with rate of fire, but not too concerned with accuracy.

An actual example from history serves to further illustrate the problems in defining effectiveness and the necessity to consider the mission in selecting MOE. In the early phases of WWII, a great many British merchant vessels were damaged or even destroyed by aircraft attacks. As a consequence, merchant vessels were equipped with antiaircraft guns and crews. After a period of time it was discovered that only 4% of the attacking enemy aircraft were actually shot down. This led some to conclude that the systems were ineffective on ships and could be better employed elsewhere, where kill rates were higher. Employing this MOE, the decision seemed inevitable. However, further examination of the data revealed that the antiaircraft fire greatly reduced the lethality of the enemy attack. In fact, the inclusion of antiaircraft weapons virtually halved the probability that a ship would be sunk. Viewed in this light, the systems were considered highly effective. In other words, the selection of the wrong MOE, or the exclusion of critical MOE, can lead to the wrong decision about effectiveness.

One further point needs to be emphasized. Training authorities and evaluators are not generally interested in the same kinds of MOE as hardware developers. The hardware is developed and fielded long before they get into the act. They must train personnel to use the equipment as it is, and must evaluate the effectiveness of the combination of the man and machine system. It matters little if a bench-fired weapon places 100 consecutive rounds within a 6-inch circle at 1000 meters, if, a typical user cannot hit a stationary enemy at 50 meters when employing the weapon. When evaluating training or unit readiness, the mission to be accomplished must be considered and the criteria of success must be set realistically in terms of the potential for man/machine effectiveness. Unfortunately, written guidance for the evaluator to aid him in selecting or developing MOE is nil.

Defining team effectiveness. One of the major problems associated with the evaluation of team effectiveness has been the inability of investigators to agree on what differentiates team and individual tasks. Most investigators agree that it is wasteful of effort to measure performance in a team context when the performance is actually nothing more than an aggregate of individual performances. Individual job skills can almost always be measured more easily, completely and cost effectively through individual job performance tests. It is felt that measurement of performance in a team context should be reserved for only those tasks which are truly team tasks; that is, *tasks which require cooperation or coordination to the extent that skills must be practiced in a team situation in order to be optimized.*

Hall and Rizzo characterized tasks performed by teams as being in either "established" or "emergent" situations. In established task situations, the sequence of task performance and the activities involved can be almost completely specified. Also, the assignment of task functions among team members and the equipment they operate are virtually fixed. In emergent situations, decision-making, problem-solving and sharing come to the forefront. The sequence of operations is not fixed, and the allocation of functions is variable. Hall and Rizzo essentially conclude that tasks performed in established situations are not really team tasks. Rather, overall task performance is simply the sum of the performances of the individual team members. Therefore, tasks performed in established situations should not be evaluated in a team context.

Unfortunately, in discussing various tasks with knowledgeable people in the armor community, I have found little agreement as to which tasks are established and which are emergent. For example, some have told me that firing on the move is definitely a team task. The advocates of this position point to the need for precise timing between the driver, who must find a level spot at exactly the right moment and maintain his direction, and the rest of the crew. Others feel that any accomplished driver does this habitually, and that so long as all crew members are individually competent, that the procedures employed ensure the proper conduct of the engagement. I will not attempt to defend either of these positions; I mentioned this example only to illustrate the differences of opinion I have encountered in trying to differentiate team performances from performances which are merely an aggregate of individual performances.

Problems with numbers. In attempting to fully describe the job situations of a tank crew in gunnery, Kraemer, Boldovici, and Boycan[8] derived a set of 11 classes of conditions or variables that could affect a crew's capability to successfully engage targets. Some examples of these classes and the number of levels identified for each class are shown in the following slide. The term "levels" refers to subclasses within a main class. If a tank gunnery objective were written for all possible combinations of levels, a total of 1,679,616 objectives would result. However, a large number of combinations are unrealistic (e.g., a moving bunker) and were discarded. Judicious combination of other levels reduced the total number of realistic combinations to the current number of 266. To test a crew's ability to perform all of these job objectives would be time-consuming, to say the least, and it must be remembered that these objectives cover only tank gunnery. Obviously, it is not feasible to measure job proficiency on all possible job objectives. Tests designed to measure effectiveness will be able to address only a limited number of the objectives. However, the need to select a limited subset of job objectives for testing is likely to produce unfortunate results. Training is almost certain to be concen-

[8] R. E. Kraemer, J. A. Boldovici, and G. G. Boycan. *Job objectives for M60A1AOS tank gunnery*, ARI Research Memorandum 76-9, Human Resources Research Organization, April 1976.

SLIDE 4

Conditions and Levels Within Conditions*

Conditions	Levels Within Conditions
Weapon	Main Gun Coaxial Machinegun Caliber .50 Machinegun
Fire Delivery Method	Battlesight (non-precision for machineguns) Precision Range Card Range Card Lay to Direct Fire
Firing Vehicle Motion	Stationary Moving
Target Visibility	Visible Without Artificial Light Visible With Artificial Light Not Visible
Target Range	<500 meters 500-900 meters <900 meters <1100 meters 1100-1600 meters 500-3200 meters 1100-2300 meters 1100-3200 meters ALL

*Condensed from FIG. 2, page 2, R. E. Kraemer, J. A. Boldovici, and G. G. Boycan, *Job objectives for M60A1AOS tank gunnery*, ARI Research Memorandum 76-9, Human Resources Research Organization, April 1976.

trated on those areas which will be tested, to the detriment of other aspects of the job. This might be avoided by testing each crew on only a small sample of jobs from the total job realm. If no crew knew exactly which set of items they would receive, they could not slant their training to the tests. However, the development of test items for every aspect of the job would be expensive. Also, the resources necessary for testing all aspects of the job would be extensive. In short, it appears that we have too many tasks and too few resources.

Reliability. We can only hope that our MOE are valid; that is, that they are indicative of how our teams would perform in combat. However, we usually can estimate their reliability. We were surprised, therefore, to find that the reliability of Table VIII scores has apparently never been determined. The only data located which even bear on the subject are those reported by Baerman and Eaton.[9] They found a correlation of r = .68 between ratings of tank commander motivation and Table VIII scores. This would indicate that the reliability of the Table VIII scores was at least 0.68. However, there were several differences between both the conduct and the scoring procedures employed by Baerman and Eaton and those typically employed. A major difference was that scoring of hits was based on a close-in, after-the-fact examination of the targets rather than by an observer riding the tank. These investigators found early in their research that the observer determinations of hits were subject to considerable error. Therefore, had the Table VIII scores been obtained in the usual manner, quite different results might have been obtained. My personal feeling is that the test/retest reliability of Table VIII scores derived as recommended in FM 17-12[10] would be unacceptably low.

Steinheiser and Snyder[11] pointed to another reliability-related problem with Table VIII. For example, assume that 70% is a passing score. Further assume that we test 100 crews whose "true" level of functioning is exactly 70%. By chance, 47 of these crews would score less than 70%, and therefore be misclassified as nonproficient. Similarly, 21% of the crews whose true level of functioning was only 60% would, by chance, be misclassified as proficient. Errors of misclassification could be reduced by increasing the length of the test to improve its reliability. However, increasing the length would also increase the resource requirements, and resources are extremely scarce at this point in our history.

---

[9] V. P. Baerman and N. K. Eaton. "Crew assignment and training," *Armor*, January-February 1977, 50-53.

[10] FM 17-12. *Tank gunnery*, HQ, Department of the Army, Washington, D.C., March 197?.

[11] F. Steinheiser, Jr., and C. W. Snyder, Jr. "Score quality issues related to individual and weapon crew criterion-referenced performance tests," presented at the Military Testing Association Conference, October 1976.

To recapitulate, our evaluations of tank crews are currently based almost entirely on performance in Table VIII. Yet, Table VIII scores are of unknown but questionable reliability. Because of this nearly total reliance on Table VIII, it is imperative that its reliability be determined, and that every attempt be made to improve its reliability, either by changes in scoring procedures or modifications to the conduct of the test. However, to date, I have been unable to obtain the necessary support to conduct a reliability study.

I have not closely examined specific team evaluation procedures in any other context. Therefore, I have no idea whether other branches in the Army or other military services face similar problems, but I strongly suspect that they do.

Testing strategies. Two principal issues divide evaluators in their approaches to testing. These are the employment of (a) one- vs. two-sided test situations, and (b) process vs. outcome measurements.

One-sided vs. two-sided tests. In a one-sided test, such as Table VIII, the examinees face a relatively structured situation in which the sequence of events is relatively fixed. "Aggressor" forces, if present at all, are restricted to specific preplanned activities. In a two-sided test, aggressor forces must be present and typically have few limitations placed on their activities. The advocates of two-sided exercises stress the importance of realism, the opportunities for real-time decision-making, and the morale-boosting aspects of competition. They also point out that the inflexibility of one-sided tests makes them easy to train and practice for. Therefore, they feel such tests provide only poor indications of how the participants would actually perform in combat.

Those favoring the one-sided approach to evaluation point to the fact that repetition of the identical circumstances is virtually impossible in a two-sided test. Therefore, no two individuals or teams receive exactly the same test, making it impossible to set exact performance standards or to compare the performance of any two teams. I should point out that choosing the type of test is not always a problem, for the type of data required frequently determine the most suitable type. For example, if exact times are needed, such as the time to fire after line-of-sight to a target is achieved, a one-sided test should be employed. Knowledge of the exact moment the target appeared would be virtually impossible in a two-sided test. One-sided tests are also generally necessary if live-fire is required.

Two-sided exercises are considered essential when targets must be generated. For example, a two-sided exercise would be necessary if the MOE were to be the ratio of friendly to threat casualties.

Process vs. outcome measurements. Stated very simplistically, "process" measurements are concerned with an evaluation of all of the actions taken during an engagement, but are not particularly concerned with

the final outcome. "Outcome" measurements are not concerned with the procedures involved or the progress of the engagement, but only in who wins and who loses.

Osborn[12] is an advocate of process measurement. He feels that to be useful, a test must be diagnostic. That is, it must provide information on exactly why a particular aspect of performance was successful or unsuccessful. Hammell, Gasteyer, and Pesch[13] state the case for process evaluations in discussing Advanced Officer (AO) tactics training as shown in the next slide. In other words, Hammell, et al. feel that process is the <u>only</u> important aspect of performance in <u>training</u> evaluations. A good decision or action may lead to a poor outcome, but the decision or action should be evaluated on its own merits, and not on the vagaries of future actions by an unpredictable enemy.

<u>SLIDE 5</u>

> ...numerous alternative sequences of actions may exist, many of which may be equally plausible for attaining a specific objective. The sequence of actions employed by the AO contains a complex series of evaluations and action selections which are situation intended. The attainment of the ultimate objective *may often be irrelevant to the evaluation of the AO's performance. This hit or miss philosophy, although distinctly meaningful in the operational environment, is inadequate in the training situation.*[14,15]

The case for outcome measurements can be stated rather simply. In an operational environment, commanders are more interested in friendly/enemy loss ratios, resources expended, and territory won or lost. The attainment of some set of predetermined mission-oriented goals along these dimensions is a much more meaningful measure of effectiveness to the field commander.

---

[12] W. C. Osborn. *Process versus product measures in performance testing*, Professional Paper 16-74, Human Resources Research Organization, Alexandria, Virginia, October 1974. (Based on paper for Military Testing Association Meeting, San Antonio, Texas, October 1973))

[13] T. J. Hammell, C. E. Gasteyer, and A. J. Pesch. *Advanced Officer tactics training device needs and performance measurement technique - Volume I*, TR:NAVTRAEQUIPCEN 72-C-0053-1, General Dynamics Corporation, Electric Boat Division, Groton, Connecticut, November 1973.

[14] *Ibid.*

[15] Italics added by author.

Perhaps you are wondering why I bring up these strategies in a paper dealing with problems. The situation as I see it is this: We need process evaluations for feedback to training managers, and we need outcome evaluations to meet the needs of field commanders. Yet, it is difficult to obtain process information from a two-sided test and even more difficult to obtain outcome information of the kind desired by commanders from a one-sided test. It is difficult enough to obtain resources for even one type of test, much less two. The problem is in finding a way to combine the best features of both types of tests without undue expenditure of scarce resources.

Resources. I have already mentioned the resource problem in passing several times. The military services are experiencing one of the longest and most severe periods of austerity in their recent history. Yet, as has been pointed out, adequate evaluations are quite demanding of resources. In less austere times, Baker and Cook[16] painstakingly constructed a "Tank Platoon Combat Readiness Check." The final checklist, including instruction to the examiner, was approximately 90 typewritten pages in length. The authors also pointed out that the entire evaluation took approximately 30 hours to administer and required the use of "aggressor" forces. At the present time, most commanders would consider the resources required for routine conduct of such an evaluation to be out of the question.

It seems obvious that we cannot develop adequate evaluation techniques for team performance unless additional resources can be found. While such is unlikely an an absolute sense, the possibility of conserving resources for evaluations offers some hope. Simulation techniques, for example, are being employed for training with increasing frequency and with little apparent loss in training effectiveness. For example, Powers, McCluskey, and Haggard[17] trained four groups of tank gunners employing 100%, 66%, 33%, and 0% live-fire. There were no differences between the hit percentages of the four groups in a live-fire posttraining test. Therefore, it appears that considerable ammunition could have been saved with no loss in training effectiveness.

Whether through the use of simulation or by other means, it is our opinion that the problem is not whether we expend the resources, but rather, how we obtain the necessary resources. As MG Gorman has stated, we must be prepared to fight outnumbered against an enemy whose weaponry

---

[16] R. A. Baker and J. G. Cook. *The development and evaluation of the tank platoon combat readiness check*, Research Memorandum, Human Resources Research Organization, Alexandria, Virginia, April 1963.

[17] T. R. Powers, M. R. McCluskey, and D. F. Haggard. *Determination of the contribution of live firing to weapons proficiency*, Final Report FR-CD(C)-75-1, Human Resources Research Organization, Alexandria, Virginia, March 1975.

will be virtually equal to ours. To do so, we must be able to accurately evaluate our fighting teams, and take corrective actions to eliminate any deficiencies.

TESTING FOR COORDINATION IN SMALL UNITS

Clay E. George

Texas Tech University

One purpose of this paper is to demonstrate the need for formal testing procedures to determine unit member response coordination in small military units. I will attempt to show that "response coordination" is a major determinant of small unit efficiency and that it is neither a "state" nor a "trait" variable. It appears to be a characteristic which an individual exhibits consistently within a given social-behavioral task setting but which may vary greatly across such settings.

A second purpose is to suggest approaches to the measurement of coordination in various types of military units. Research efforts over the past 20 years have made it possible to specify precise measurement operations for the rifle squad. They have also produced guidelines for establishing such operations relative to other types of units.

A model of unit performance (Figure 1)

My analysis of the determinants of performance in mission oriented groups is probably very much like any one else's. A group can not perform well if its members are not competent. The contributions of psychologists and others have been very great in establishing effective procedures for individual selection and training in military environments. In effect, we have some rather impressive technologies for ensuring that our units are composed of proficient individuals, within the limits of available human resources.

Higher level leadership and "organizational climate" also affect unit performance. I think that the ideas presented here are also relevant to this topic (Klein, Gibbs, George, Pruitt & Patrizi, 1977) but my central focus is on the lower level unit.

A number of "morale" variables have been shown to correlate with unit performance, especially in stressful circumstances (e.g. Dudek, George & Ayoub, 1969; George, 1965; George, 1970 and Hodge, 1972). The weight of the evidence, however, seems to indicate that these morale variables are often symptoms or effects rather than "causes" of performance.

The primary concern here is with delineating, observing and counting specific coordinative responses available to unit members. To fur-

ther this objective, type of unit is roughly categorized in a second model.

Model of unit types (Figure 2)

It has been necessary to take into account both the degree of group structure and the flexibility of that structure to explain the findings from my research program. Degree of structure is measurable in several ways (George, 1962). Perhaps the simplest way for present purposes, is to use the ratio between the number of role specialties and the number of group members. If each unit member has a unique specialty within the group, that group is completely (100%) structured. If every unit member could have exactly the same role specialty (no leadership), the group would be completely unstructured. Military units tend to be relatively highly structured. We are, therefore, primarily concerned with the higher end of this dimension.

Flexibility of structure is measured (or estimated) by the probability of role interchange in the operational (not necessarily training) environment. The rifleman in an infantry squad has a very high probability of being required to take over some (or all) of the roles of a grenadier or team leader during operations. A steward on a MATs flight, on the other hand, has a low probability of taking over the pilot's role successfully. Highly but flexibly structured small units will be called teams. These include army and Marine Corps infantry squads and Navy and Air Force advanced base security units, among others. Crews are highly, but less flexibly structured units such as aircraft and tank crews. The team-crew distinction is a useful aid to the recognition and measurement of intraunit response coordination.

Measurement Approaches (Figure 3)

Symptomatic variables. Cohesion/status is perhaps the most thoroughly studied of this class of variables. A common measurement process involves rankings or ratings by each group member of the respect or affection they hold for each other member. This measurable aspect of units does tend to correlate with performance and it probably helps to protect performance levels from the deleterious effects of stress (George, 1962, 1965, 1967, 1970). The validity exhibited by cohesion as measured in research settings, however, tends to wane if used administratively in ways that might produce contingencies for the group or for any member(s) thereof. The use of peer ratings or rankings in officer candidate programs produced fairly good prediction of combat performance in WW II, a lower level of prediction during the Korean conflict (cf. Jennings, Rose & Kreug, 1974) and a negative relationship with leadership knowledge and skills during the Vietnam era. The latter finding is based on unpublished research undertaken by myself and others as a part of technical advisory

services during the late 1960s. Any administrative use of this sort of measurement process should be undertaken with great caution.

A second set of symptomatic variables results from Carter's (1955) derivation of three factors describing the behaviors of persons in small groups. Inventory measures have been reported by Bass (1962) and by George (1967). These scales indicate the person's tendency in groups to maximize: (1) personal achievement (intragroup competition), (2) socializing and/or (3) unit efficiency (coordination). Unit member scores on the motive to coordinate do predict unit performance on tasks requiring coordination, but those who start out low on this characteristic show significant increases when their unit is reinforced for improved performance (George, Hoak & Boutwell, 1963; George, 1967).

Each of the symptomatic variables is useful as a research tool when looking for behavioral measures of unit functioning. Using these variables as primary predictors or indicants of unit characteristics is a complex and uncertain procedure, however. One is well advised to proceed with caution.

Behavioral response coordination

At the most general level, coordination is composed of recognizing and acting upon a unit's need without specific direction or instruction to do so (Figure 4). This level may be illustrated by a study of 51 Army ROTC sophomore cadets working in 8 "leaderless" groups with 4-9 members each (George, Simms & Lumpkin, 1969; George, Simms, Deardorff & Hafer, 1969; George & Dudek, 1974). The groups were given the task of assembling, from a stack of parts, one fewer rifles than the group had members. Those who quickly assembled a rifle, or who had no parts to work with, could choose to help others or to stand idle during the work period. One observer per cadet tallied each unrequested, spontaneous action (suggestion, direct aid, etc.). After the task was completed, each cadet wrote a critical incident report on at least one other cadet whose behavior had effected the group's performance. Spontaneous coordination (initiative) correlated .47 with number of positive critical incidents credited by peers, -.37 with negative critical incidents and .45 with global peer leadership evaluations taken approximately 6 weeks later. Each of these correlations was significant at the .05 level. Grade point average had an insignificant correlation of .04 with coordination. The general pattern of results indicates that, even in low structure groups, one can measure coordination with some degree of consensual validity.

Experiments with intact fire teams and rifle squads (highly structured units with missions requiring flexibility of structure) also show that observers can count coordinative responses (George, 1967, 1970). Coordination responses in fire teams acting as base of fire elements in-

cluded increasing sector of fire when another could not cover his sector, redistributing ammunition as required, taking the position of leader or automatic rifleman when casualties occurred, etc. This class of responses increased over training trials (each trial facing the team with different emergencies requiring different specific responses from the various people involved) from 21% of the needed responses on trial 1 to 65% on trial 4 ($p < .05$). Over the entire problem, coordination scores correlated .72 with a criterion fire distribution score and .56 with a criterion fire volume score. Both correlations were significant at the .05 level. In team situations, response coordination can be measured, it can be increased by appropriate training and it can be shown to improve performance on criteria of military importance.

The specification and measurement of spontaneous intracrew coordination is, unfortunately, less well established. It is true, of course, that crew coordination can more often be achieved by leader direction and control or by machine provided cues than is the case of teams (George, 1970; Miller, 1971). Still, crews with military missions will face situations requiring spontaneous coordination and outcomes may be catastrophic in its absence. Task relevant communications are most often suggested as appropriate measures for crew coordination (Brown, 1977; McRae, 1966). A major problem is that efficient crews operate with an absolute minimum of communication of any kind. There are indications in the literature that number of task relevant communications may correlate positively with performance given tasks of sufficient difficulty (Figure 5).

McRae (1966) studied 12 four-man crews of soldiers solving problems of increasing difficulty over three trials. After partialing out solution time, significant correlations of .68, .76 and .88 were reported between task relevant communications and response accuracy. The more difficult the problem, the more communication served a coordinative function to enhance performance. Although this explanation is not a compelling one from McRae's data there are additional experiments to support it.

Brown (1977) studied 12 four-woman teams working problems under each of four conditions of difficulty level. Time was held constant. Correlations between task relevant communication (requesting information) and number of correct solutions was -.89 for crews working on the easiest problems and -.31, -.16 and .19 for those working on progressively more difficult problems. Although only the largest of these correlations is significant, it does appear that sufficiently difficult tasks may force crew members to communicate when, and only when, task requirements demand such coordinating responses.

George, Keating, Lumpkin and Miller (1971) reported that five-man crews could perform well with very limited communication allowed to them

provided that they had been given an earlier training trial with unrestricted communication opportunities. Conversely, teams trained under severely limited communication conditions failed to perform well on even the second of two transfer trials conducted with unrestricted opportunity to communicate. Perhaps we will have to learn the specifics of coordination by studying crews under very difficult task conditions.

Summary

Learning to measure coordination in the rifle squad has provided a general approach (Figure 3) which should be applicable to military teams in general. Indications are that a research effort similar to that used with teams will produce equally good measures of crew performance; that is, study crew performance under very difficult circumstances (Figure 5). Small unit evaluation is believed to be an application of behavioral science which could be of major value to the military services. Such measurement should greatly effect unit training programs and readiness evaluation.

References

Bass, B.M. The orientation inventory. Consulting psychologists press, inc., Palo Alto, Ca., 1962.

Brown, P.E. The effects of task difficulty on performance in female cooperative groups. Unpublished master's thesis, Texas Tech University, 1977.

Carter, L.F. Recording and evaluating the performance of individuals as members of small groups. In: Small groups. Hare, A.P., Borgatta, E.F. and Bales, R.F. (Eds), Knopf, New York, 1955.

Dudek, R.A., George, C.E. and Ayoub, M.M. Performance, recovery and man/machine effectiveness. Proceedings: US Army human research & development conference, 1969, 15, 177-188, AD714375.

George, C.E. Unit confidence building via teamwork training. Paper presented, HumRRO confidence building--stress conference. Alexandria, Va., 1965.

George, C.E. Training for coordination within rifle squads, HumRRO, Alexandria, Va., 1967, AD-653845. Also published in; Proceedings: US Army human factors research & development conference, 1966, 12, 80-86.

George, C.E. Goal directed leadership: Superordinate to human relations. Professional paper, HumRRO, Alexandria, Va., 1967, AD-649864.

George, C.E., Hoak, G.R. and Boutwell, J. Pilot studies of small group effectiveness. Research memorandum, HumRRO Division No. 4., Fort Benning, Ga., 1963, AD-627214.

George, C.E. Specification and measurement of intragroup coordination in various types of tasks and work groups. Technical memorandum 19-70. Human engineering laboratories, Aberdeen, Md., 1970.

George, C.E., Simus, J.D. and Lumpkin, M.A. Education for responsibility. Texas Tech University, 1969, AD-703844.

George, C.E., Simus, J.D., Deardorff, M. and Hafer, M. Education for responsibility, part 2. Texas Tech University, 1969, AD-703845.

George, C.E., Keating, P.H., Lumpkin, M.A. and Miller, D.I. Communication and group performance. Journal Supplement Abstract Service, 1971, MS 39, AD-729037.

Hodge, D.C. (Ed) Military requirements for research on continuous operations. Technical memorandum, Human engineering laboratory, Aberdeen, Md., 1972, AD-744782.

Jennings, J.R., Rose, R.M. and Kreuz, L. Stress and performance before and after officer candidate school. Journal of applied psychology, 1974, 59, 500-503.

Klein, G., Gibbs, M., George, C., Pruitt, B. and Patrizi, F. The effect of organizational structure and flexibility on performance and satisfaction. Paper presented; Southwestern psychological association, Ft. Worth, Tx., April 1977.

McRae, A.V. Interaction content and group effectiveness. Technical report, HumRRO, Alexandria, Va., 1966, AD-637311.

Miller, D.I. Skill dilution and skill level requirements as determinants of crew performance. Unpublished doctor's dissertation, Texas Tech University, 1971, AD-734182.

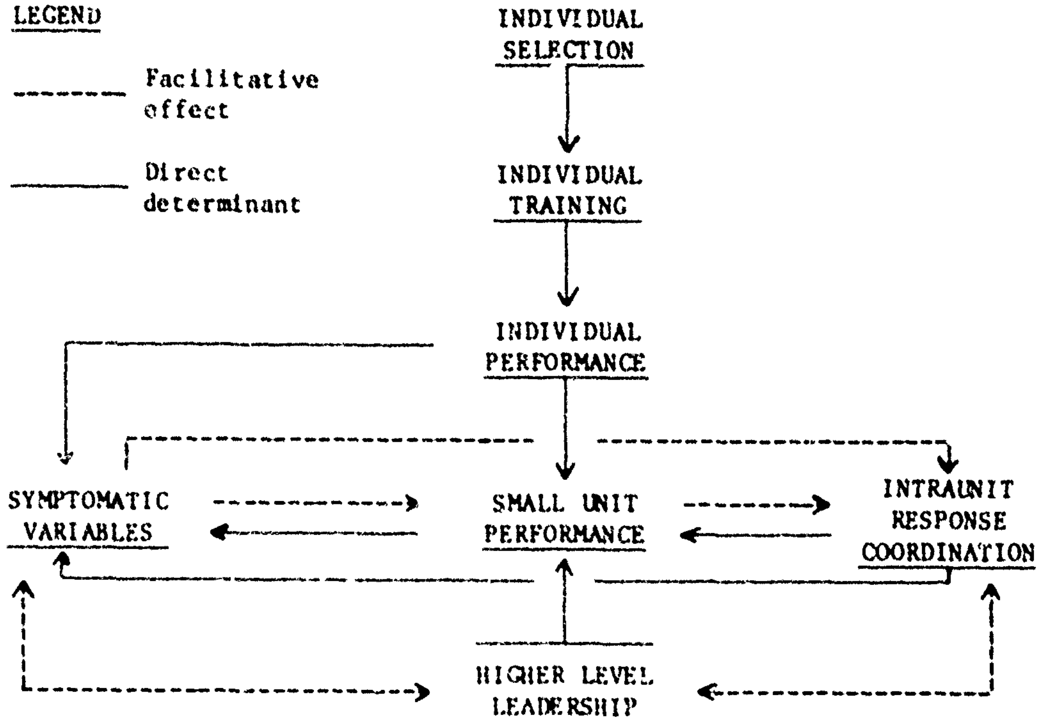
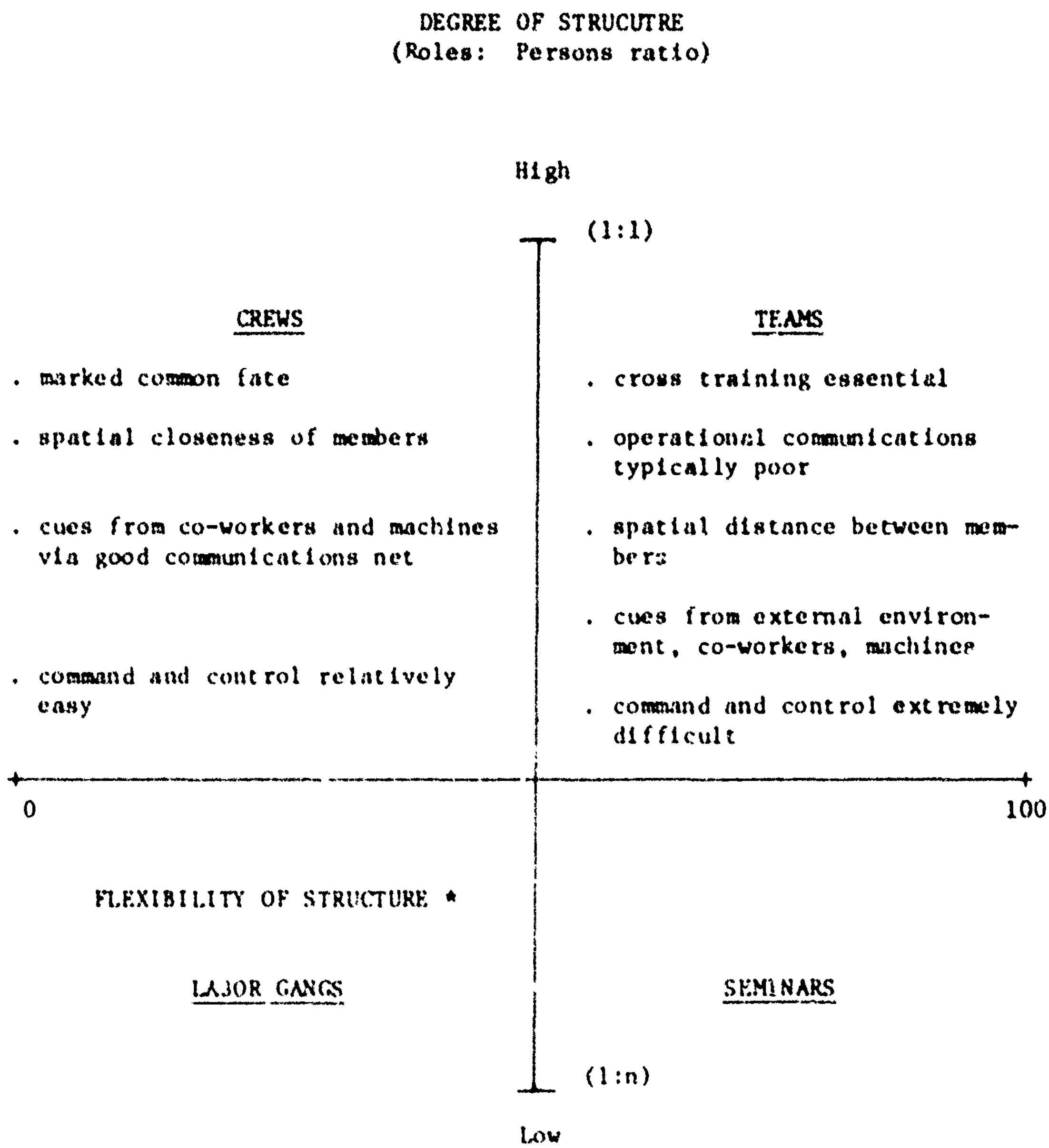


Figure 1. A model of small unit functioning.



*Probability of role interchange forced by uncontrollable events.

Figure 2. Model of small unit structural characteristics.

I. Symptomatic variables (individual and group characteristics within the unit-task-setting environment)
   A. Sociometric (questionable administrative utility)
      1. affection (stress resistence)
      2. respect (mutual confidence)
   B. Unit member motivation to maximize:
      1. personal achievement (intragroup competitive)
      2. socializing (emotional support)
      3. unit efficiency (coordination)

II. Behavioral coordination of response
   A. Shared attention among:
      1. one's primary job
      2. status of co-workers
      3. machine(s) in the unit system
      4. extra-unit task enviroament
   B. Recognition of initiative taking requirement
   C. Respond to requirement
      1. individual, immediate action
      2. communicate status to other(s)

Figure 3. Small unit level correlates of performance.

Coordination required by:	Cue-event:	Coordination requirement:	Cost if uncoordinated:
Enemy action, weather, terrain	Lost effectiveness, person/weapon	Take over, inform others	Poor fire volume and distribution
Same as above	Loss of contact	Regain by movement, voice. visual search	Decreased effectiveness
Fear, terrain, poor visibility	Inadequate dispersion	Moving, getting others to move, encouraging	Unnecessary exposure and loss
Heavy or surprise fires	Going to ground	Initiate fire and movement even without orders	Destruction of unit
Personnel turnover	Receipt of replacement	Accepting, supporting emotionally, training	Skill dilution, lost cohesion
Combat stresses	Indecisive behaviors	Suggesting, encouraging, correcting	Loss of unit initiative, drive

Figure 4. Operational conditions leading to coordination requirements by team members.

Source: Small unit combat after action reports in Infantry School library, Fort Benning, Ga.

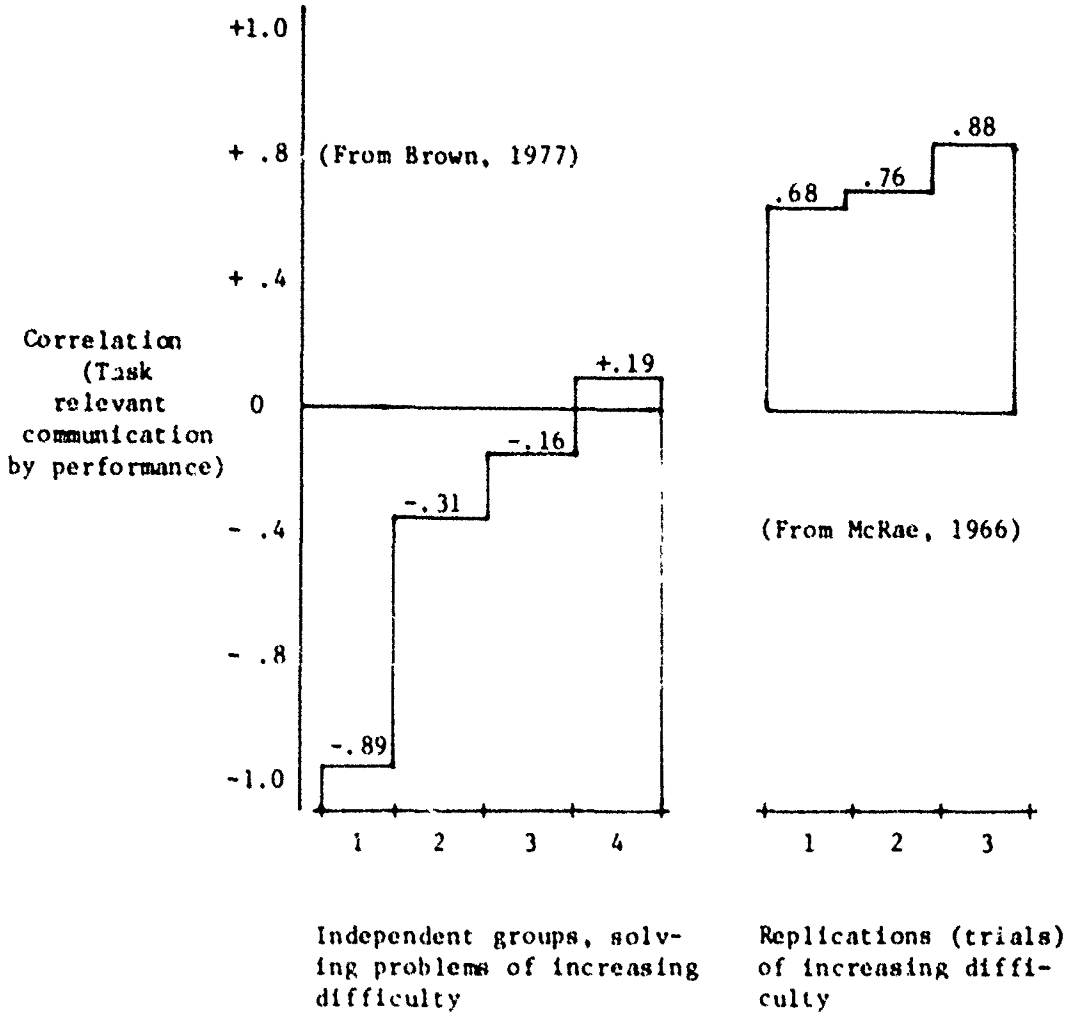


Figure 5. Task oriented communications in crews as correlates of performance mediated by task difficulty.

A CULTURE-FREE PERFORMANCE TEST OF LEARNING APTITUDE[1]

James K. Arima
Naval Postgraduate School

From World War I to the late 1950s, standardized mental tests with nationally based norms became widely used for selection, placement, and classification decisions. Their great acceptance was due, in large part, to their role in furthering the American concept of an egalitarian society (Holzman, 1971). That is, decisions of considerable importance to individuals could be made on the basis of merit, given a person's score on an objective test of ability with the requisite reliability and validity.

The Armed Services were leaders in the testing movement, and the use of the Army Alpha and Beta tests in World War I has been identified with the beginning of the testing movement in which large numbers of persons are routinely tested for selection and placement. Nearly two million people were given the tests during the course of the war, and the results provided much of the information for later studies of demographic, socioeconomic, and cultural differences in intelligence and ability (Matarazzo, 1972). World War II saw a similar emphasis on mass testing and the development of the Army General Classification Test (Melton, 1957). Again, the results of the testing program provided large amounts of valuable information for scientific study that went far beyond the limited purposes for which tests were originally administered. Eventually, the AGCT was made available in commercial form for sale to qualified users in the general public.

In the post-World War II years, the Armed Forces Qualification Test (AFQT) with a scoring in readily understandable percentiles became the standard, general test of mental ability for the services. The AFQT designation of mental categories is still in use today. Throughout these developments, sepcial-purpose tests were also being created by the individual services until a common entrance test was no longer the rule with the advent of the All Volunteer Force (Melton, 1957; Windle and Vallance, 1964). More recently, however, an emphasis on efficiency in the testing program on the part of Congress and the Defense Secretariat has seen the emergence of the Armed Services Vocational Aptitude Battery (ASVAB) as a common test of general aptitude for military service. A form of the ASVAB is also used in civilian, secondary schools in the High School Testing Program managed by the Armed Forces Vocational Testing Group (AFVTG).

---

[1] I am indebted to Peter A. Young for running the subjects and collecting and analyzing the data as a part of his master's thesis (Young, 1975). Paul Sparks created the instrumentation for the experimental administration of the test. The terms culture-free and culture-fair will be used to mean the same thing.

The growth and apparent success of the testing movement has not been without its critics and detractors. The criticism did not reach social significance until the middle and late sixties when many of our institutions were put to severe test with a reexamination of our value systems and the emergence of new concepts for improving the quality of life in America. The routine testing of job applicants took a severe setback in the Griggs et al. vs. Duke Power Company decision of the U.S. Supreme Court when it ruled that a test could not be used as a selection device unless the measured abilities represented by the scores on the test were shown to be required for acceptable performance on the job. This decision had at least two implications for testing. One, obviously, related to the traditional concept of the predictive validity of tests, and the other was with respect to the use made of tests.

Regarding the predictive validity of tests, the court's decision was quite telling, since most tests predict intermediate criteria well--such as normatively scored achievement tests--but not more distant, more ultimate criteria, such as occupational success (Goslin, 1968). This situation is particularly prevalent in such large institutions as the military (Thomas, 1972a, 1972b) and the nation's educational systems. The question of the use, or misuse, of tests focuses on the results that testing programs produce. The argument has been that differential prediction or classification of individuals results when they are categorized on the basis of ethnic and socioeconomic backgrounds. Broadly stated, differential prediction means that the proportion of individuals who, for example, pass a selection cutoff score is not the same for the different categorical groups. Such differential prediction has been labeled bias because culturally deprived persons have not had the opportunity to master the material content of the tests nor to develop the test-taking motivation, experience, and specific skills of other groups of persons (Goslin, 1968). The bias is usually attributed to the test, rather than to the uses made of the test, but the argument is not entirely convincing (Green, 1975). Even on a strictly psychometric basis, several different definitions of bias are possible (Hunter, Schmidt, and Rauschenberger, 1977).

While the Armed Services have managed to escape severe criticism in the past, there are signs that the situation is changing. The use of the ASVAB in the High School Testing Program recently received very sharp criticism from Lee J. Cronbach, and the Office of Management and Budget (OMB) has instituted a series of inquiries into the management of their testing programs on the part of the several services.

Complicating the issues of test validity and test usage as sources of bias is the argument with respect to the roles of heredity and environment in the determination of a measured, mental ability--such as intelligence. If, as argued by Jensen (1968a), heredity plays the predominant role by a margin of as much as 2-to-1, then the cultural deprivation argument loses considerable weight. That is, the important

differences exist, more or less, independent of environmental factors. On the other hand, if it is argued that the range of performance capabilities at a fixed hereditary level is broad and essentially unpredictable due to the influence of many environomental factors (Feldman and Lewontin, 1975), then the role of cultural and socioeconomic factors in causing the differential prediction of testing programs must be acknowledged and corrected. A deceptively simple solution would be to create tests that are culture free. Presumably, a culture-free test would be measuring the "real" or hereditary potential--the genotype--of the person being tested. But, if an operational definition of an unbiased, culture-free test is that all categories of cultural groups have the same mean and distribution function on the test, the use of such a test for selection is highly likely to result in differential outcomes on some criterion measure, such as the ability to complete a course of training within a prescribed or reasonable period of time. The test has been made culture free, but it has little or no predictive validity. The argument could be made that the fault lies in the criterion, and not the test. In this case, a third fundamental question regarding the testing movement arises, and that is the construct validity of a test or what is the test supposed to be measuring? (Goslin, 1968).

As explained in the preceding argument, the creation of a culture-free test places a greater burden on the construct validity of the test rather than its predictive validity, since it may not be possible to determine the latter in the traditional manner. In addition to escaping criticism for being biased, a culture-free test of mental ability with high construct validity would be of great value to the military services and other large institutions that face increasingly difficult problems in personnel procurement owing to the shrinking of the pool from which new recruits must be obtained (Congressional Budget Office, 1977). Under these circumstances, if standards are not to be lowered, means must be found to identify individuals with high native ability who do not score well on traditional tests. It was the purpose of this project to explore the possiblity of developing such a test that was relatively culture-free, had high construct validity with respect to identifying individuals of high native ability,and would be feasible and practical to administer in the military testing environment.

TEST DEVELOPMENT

THE MODEL

The first problem in developing the test was to find a model upon which to build the test. A model, in this usage, is a procedure or paradigm that reliably elicits for quantitative measurement a behavior that is the result of a cognitive process that is frequently involved in many situations in real life. Models of this sort would be available in such traditional experimental areas as learning and memory, information processing, problem solving, and decision making. It was felt that most of the paradigms for information processing placed an overly high

emphasis on verbal behavior and materials and that this feature would make it difficult to achieve a culture-free test. The problem-solving paradigm was thought to be inappropriate for test construction from a reliability and measurement standpoint, since an attempt to control and standardize the set or approach an individual takes would tend to destroy the objectives of the paradigm, itself, which encourages experimentation by the subject. Also, the frequency of chance or "a ha" solutions would tend to make test scoring difficult, categorical, and unreliable. The decision-making paradigm was not considered appropriate because of the paradigm's reliance on value systems in the elicited behavior--value systems developed through life experiences and very much the product of an individual's culture.

This left the area of learning as a logical choice for the model. Learning paradigms have been the traditional vehicle of the majority of research in the behavioristic tradition, and learning ability is generally recognized as an important ingredient in an individual's adaptation to a job. In the industrial engineer's armamentarium, the "learning curve" is an important ingredient for an entire production process. There are many reliable measures of the learning process--at least in the aggregate. And the law of effect, in its empirical form, is without precedence among the many, so-called "laws" in psychology. As quoted and discussed by Estes (1974), Thorndike believed that intellect is the ability to learn and that estimates of intellect should be estimates of the ability to learn. In another sense, Thorndike believed that intellect is the ability to learn more things or to learn the same things more quickly. Typical intelligence tests that sample the products an individual is able to produce seem to be assessing intelligence with respect to the amount of stored information, knowledge, and intellectual skills, whereas the typical experimental learning paradigm would seem to consider the rate of learning as a measure of intellectual performance.

Within the field of learning, visual discrimination learning was selected as the general paradigm in which to build the test because it has been widely used at many phylogenetic levels to study the evolution of intelligence (Bitterman, 1965, 1975). There is also an extensive literature in the visual discrimination learning of human subjects as well (Green and O'Connell, 1969). The typical paradigm for visual discrimination learning involves two or more dissimilar, visual stimuli of which one has been arbitratily designated as correct. The organism learns to respond to the correct alternative--e.g., peck the middle disc--by being reinforced for making the correct choice.

Examination of the Green and O'Connell (1969) bibliography will show that most of the experimental tasks in visual discrimination learning have been relatively simple owing to the design of such tasks for animals, children, and retardates. The visual discrimination learning situation has been made more complex by manipulating reinforcement contingencies or the quality of reinforcements. In their altered form, emphasis has been on such phenomena as reversal learning, probability learning, and the effects of partial reinforcement and incentive contrasts. Bitterman has shown that the acquisition (learning) curve may

be very similar for all organisms, but the switch to one of the other conditions following original learning has led to qualitatively different behaviors by different species. Thus, it would be highly desirable to adhere to the basic learning paradigm but make the task more demanding for the human subject. This could be done by having an individual learn several discriminations simultaneously, which shall be called multiple discrimination learning. Except for the fact that pictorial materials would be used, the situation would be very similar to verbal discrimination learning (Eckert and Kanak, 1972). In a typical verbal discrimination learning experiment, a list of several word pairs is created in which one member of each pair has been designated as the correct alternative. The pairs, referred to as items, are presented individually and a complete presentation of the list is a trial. The subject instrumentally learns the correct alternatives by being reinforced when the correct member of the word pair is vocalized. Arima (1974) has shown that the paradigm is very robust in the sense that the learning rate is constant regardless of the number of alternatives (up to four) presented in a stimulus (item) as long as the information presentation rate is also constant. The key to determining this relationship was the measurement of information content in terms of Shannon bits and learning in terms of the information transmission rate.

To recapitulate, the model for the test was a visual discrimination paradigm presented in the manner of verbal discrimination learning experiments. That is, the model calls for the subject to learn several visual discriminations simultaneously, a process that will be referred to as multiple discrimination learning.

STIMULUS MATERIALS

Construction of a multiple discrimination learning test required a relatively large set of stimuli that were homogeneous, yet discriminable, and which were as free of cultural influence or implications as possible. Homogeneity of stimulus materials was desired so that each of the stimulus pairs within a "list" could be of comparable difficulty and so that any stimulus pair would be representative of the test task. Geometric shapes were eliminated because of their limited numbers and the possibility that their familiarity and association values might be linked with cultural variables. Color, hue, and brightness were also rejected because of the difficulty in production and replication and because difficulties in sensory discrimination might result when a large number of items was required. Additionally, there would be the problem of using the test with colorblind individuals. For these reasons, two-dimensional, black-and-white patterns of uniform size were investigated. The set of 30, two-dimensional, random-shaped, metric polygons used by Arnoult (1956) was found to fit the requirements admirably. They are shown in Figure 1. Moreover, they had already been categorized, as a group, as figures having high discriminability.

Prior to constructing pairs and lists of items using the forms, it was necessary to obtain measures of the pairwise similarity of the

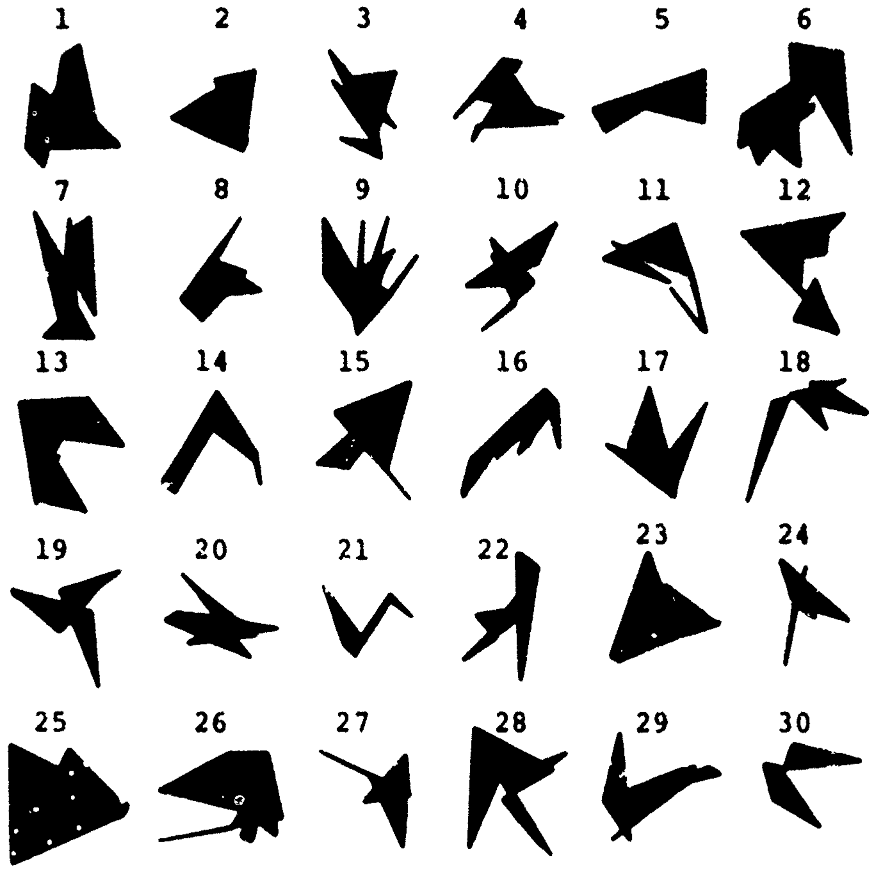


FIGURE 1. Shapes selected for use in assembling stimulus lists.

(From Arnoult, 1956)

forms and to develop a set of pairs for which there would be assurance that either member would be likely to be chosen as a correct alternative on a first (guess) trial. It was particularly necessary to develop pairs with an a priori choice of 50-50 for either member so that the information content (uncertainty) of each item would be at a maximum (1 bit) and constant within all lists. The similarity measure was desired because similarity had been found to be a significant variable affecting learning rate in verbal learning under some conditions. Accordingly, it was assumed that similarity among and between the stimuli should be controlled in constructing the test items.

In order to obtain empirical values for these relationships among the forms, a small, data-gathering experiment was conducted. The 30 stimulus polygons were arranged in pairs. All possible pairs were constructed under the constraint that an item would not be paired with itself. Left-right order within a given pair was not considered. This resulted in the assembly of (30 x 29)/2 = 435 different pairings. These pairs were then arranged in three columns on sheets. Three separate booklets, each containing 145 pairs, were constructed and distributed to 60 graduate students at the Naval Postgraduate School. Each subject received a single booklet selected at random from the three, and was asked to perform two separate tasks--selection of one item from each pair and rating of the degree of similarity seen between the items of each pair. Subjects were told that one item in each pair had been arbitrarily designated as "correct," i.e., the desired response, and were asked to designate that item which they thought to be the "correct" response. This selection was to be made with the knowledge that designation of the "correct" response was made completely arbitrarily.

Subjects were cautioned to make their choices solely on the basis of a given pair alone, and without regard to previous selections. This exercise was intended to simulate as closely as possible the condition of facing a stimulus pair in a forced-choice situation with no prior knowledge of the correct item in the pair.

Subjects then went through the list a second time, rating each pair as to whether the two items in each appeared to be very similar, slightly similar, or dissimilar. Each pair was then assigned a similarity factor of one, two, or three, respectively.

The choice preferences of the 60 subjects (20 for each set of 145 pairs) were translated into percentages and cast into a matrix. In addition, averages of similarity ratings given for each pair were computed and cast into the same matrix format. Thus pairwise estimates of choice preference and item similarity were obtained and placed in usable form.

CONSTRUCTION OF TEST LISTS

A subgroup of pairs was selected from the original 435 that had been rated. These pairs were singled out on the basis of choice preference. Subjects making choices within these pairs had displayed no

significant preference, on the average, for either item in each pair (selections were distributed either 50%-50% or 45%-55% between each). This subgroup was then used to construct the test lists. Since no marked preference for a given item in a pair had been demonstrated, it was felt that the choice probabilities associated with each could be considered to be "equally likely" for the purposes of evaluating the information content of the choice associated with each pair.

Three stimulus lists of six pairs each were constructed from the "equally likely" subgroup of pairs. These lists were assembled under the following constraints with respect to the similarity variable:

List I. Figures in each pair were as dissimilar as possible. In addition, all figures in the entire list were as dissimilar as possible. (Within-pair similarity factors were at least 2.50, averaging 2.60, while between-pair factors were not less than 1.75, averaging 1.98.)

List II. Figures in each pair were as similar as possible, but dissimilarity between pairs was maintained. (Within-pair similarity factors were no greater than 1.95, averaging 1.58; the between-pair factors were no less than 1.90, averaging 2.20.)

List III. Figures were as similar as possible, both within each pair and between other figures in the list. (Within pair similarity factor was no more than 1.90, averaging 1.73; between-pair factor was no greater than 2.30, averaging 1.92.)

These lists are presented in Figures 2, 3, and 4, respectively. As can be seen, the lists were constructed in order to present discrimination tasks of increasing difficulty. Stimulus items in List I were chosen to be as distinguishable as possible, minimizing intra- and interpair confusion. Similarity within pairs was added in List II, but each pair was kept as distinguishable as possible from other pairs in the list. Similarity was extended to cover all items in List III. List III, of course, is the most homogeneous.

When lists of six pairs each had been completed, three test lists of 60 pairs were assembled. Each test list consisted of 10 repetitions of each of the six pairs of Lists I, II, and III. Order within these replicates was random. Left-right order within pairs was varied in a random fashion as well with the restriction that a given form was seen on the right five times and on the left five times. At least one different pair was presented before a given pair was repeated. The polygons were not rotated or reversed, but were presented "upright" at all times.

Thus each test subject could be presented a total of 60 pairs of stimuli. Pairs appeared in no apparent order, and the correct response was not always on either the right or left side; subjects were forced to learn the correct response in each pair solely on the basis of recognition of the items within that pair alone.

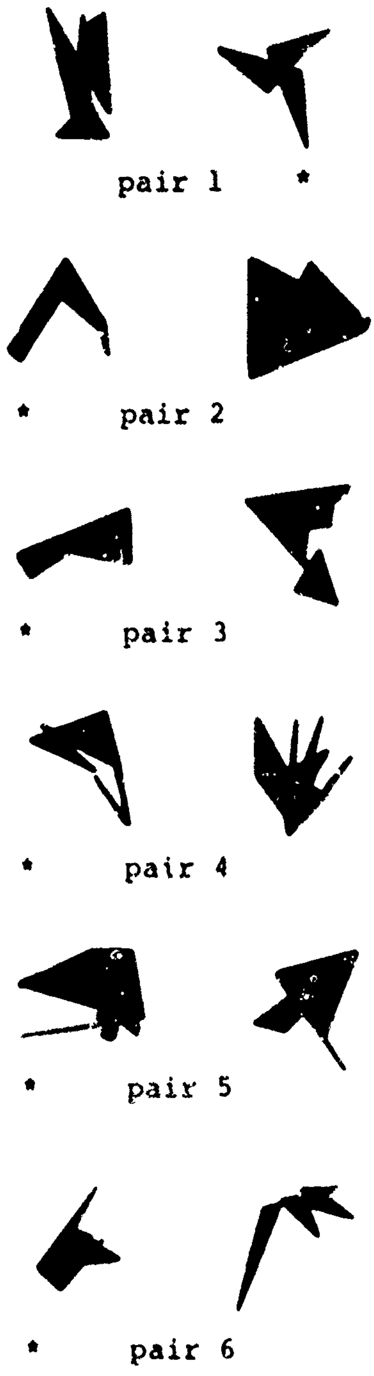


FIGURE 2. Stimulus List I.
(Least similarity within and between pairs)
*Indicates "correct" shape

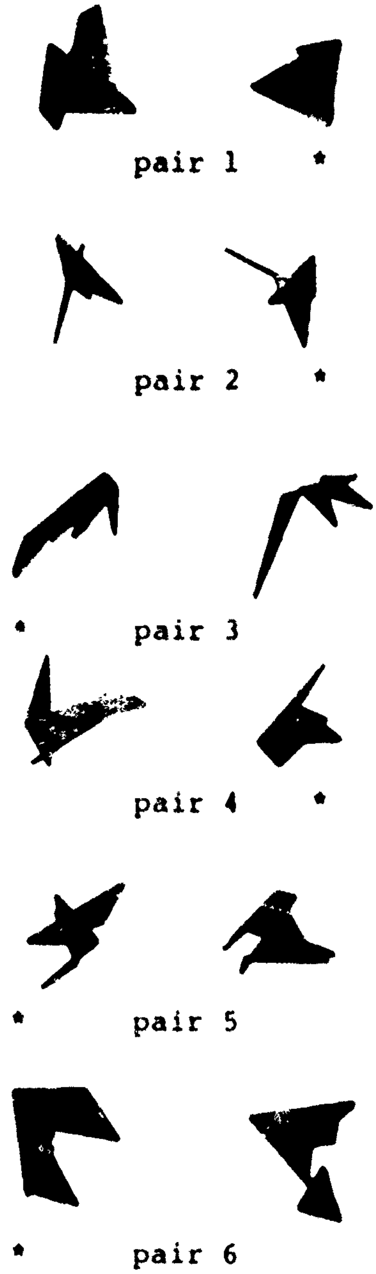


FIGURE 3. Stimulus List II.

(Maximum similarity within pairs; minimum similarity between pairs.)

*Indicates "correct" shape.

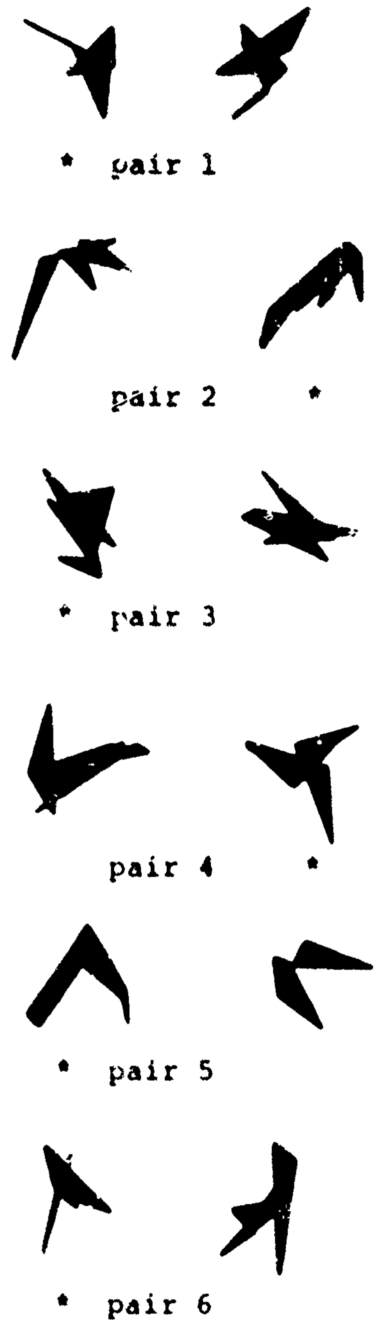


FIGURE 4. Stimulus List III.
(Maximum similarity both within and between pairs.)
*Indicates "correct" shape.

TEST APPARATUS

Test apparatus was designed to provide maximum flexibility in test administration. The apparatus array used in administering the test is diagramed in Figure 5. Critical units of the presentation and response equipment were secured in place throughout the course of test administration. Distance from the subject (edge of table) to the viewing screen was 42.5 inches (107.95 cm); reinforcement lights were located 8.5 inches (21.59 cm) in front of the screen. Stimulus pairs occupied an area on the screen approximately 6 inches (15.24 cm) high by 9 inches (22.86 cm) wide.

Stimulus pairs were mounted on 35 mm slides, one pair to a slide. Since each list was presented a total of 10 times, the 60 slides required for each list were placed in a carousel. Stimuli were rear projected onto a Kodak shadow-box screen using a Kodak Ektographic Carousel slide projector. A neutral light-reduction filter (Kodak Wratten gelatin filter, no. 96 ND 0.50), rated to reduce light transmission by 50 percent, was fixed over the projector lens to reduce excessive glare on the screen.

A modified Ohr-tronics eight-channel paper-tape reader was used to control the reinforcement lights (described below) so that only correct responses would receive reinforcement. Wiring was accomplished so that the pulse used to advance the slide projector to the next stimulus pair also advanced the tape reader. Tapes were punched to coordinate with the ordering of the stimulus list in use.

The apparatus was designed to permit a machine- or self-paced mode of presentation. Stimulus presentation rate in the machine-paced mode was controlled by an interval timer. The timer was set to provide an actuating pulse to both projector and tape reader simultaneously every 4.0 seconds. The time required for the slide projector to cycle from a presented slide to the next slide was found to be 1.0 second. Since the projection screen was blank during this cycle time, the stimulus pairs were visible for only 3.0 seconds before the timer initiated the next sequence.

Stimulus presentation during the self-paced mode was controlled by either of two identical buttons located on the sides of the response box. Pressing either of these buttons initiated the electrical pulse that advanced the slide projector and tape reader. (These buttons were inactivated during the machine-paced mode to preclude accidental disruption of the stimulus presentation rate.)

Two identical buttons fixed on top of the response box were used to designate choices. Correct responses were reinforced by one of a pair of 2.5 watt lights placed on a small box directly in front of the viewing screen. Incorrect responses received no reinforcement. Responses, regardless of reinforcement, were recorded on a two-channel Clevite brush recorder. The tapes thus obtained could be used to confirm observed responses, and in the self-paced mode to measure inter-response time and total test time.

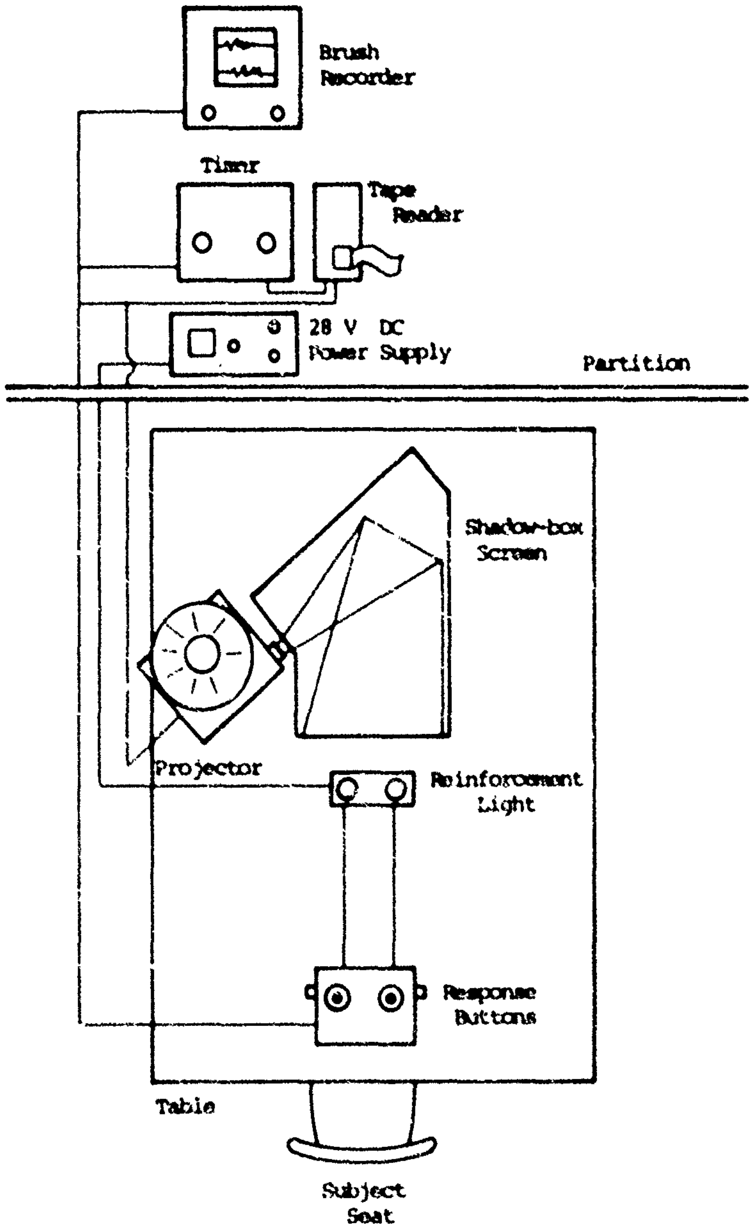


FIGURE 5. Layout of Test Equipment

Twenty-eight volt DC current to power the tape reader and reinforcement lights was obtained from a Power Designs, Inc., Model 3650-S DC Power Supply.

TRIAL ADMINISTRATION

In order to evaluate the characteristics of the constructed test under conditions as close to operational as possible, and also to investigate the appropriateness of the various test parameters (list length, similarity, etc.), it was decided to administer the test to as many subjects as possible during a five-week period in which they would be available.

METHOD

Facilities

Testing was conducted at the Naval Training Center (NTC), San Diego, California. All testing was performed in an isolated room at the Personnel Testing and Classification Center of the NTC. Since activity was planned for both morning and afternoon periods, windows in the testing room were covered with opaque material to reduce anticipated glare from sunlight and to achieve uniform lighting conditions in the room.

Subjects

Subjects tested were 160 male U.S. Navy recruits at NTC. Ages ranged from 17 to 26 years, with the average being 19 years. Average stated schooling level for the group was 12th grade (11.78). Schooling level within the nonwhite subgroup was slightly higher (12.2 years) than the group average. Nonwhite subjects were predominantly Negro, although the sample contained Oriental, Malay (Filipino), and Mexican-American recruits as well. Subjects were assigned to the various test conditions in order of appearance.

Test Design

The experiment was conducted using four test groups. Forty-four subjects were given the test using self-pacing to control the stimulus presentation rate. Test List I was used throughout the self-paced phase. The remaining three groups used the machine-paced mode to present the stimulus pairs at a constant rate of one each 4 seconds. In the three machine-paced phases, 43, 40, and 33 subjects were tested using Test Lists I, II, and III, respectively. Tabular representation of this test design is shown in Table 1. There it can be seen that the test variables were pacing mode (self- and machine-paced) and test list, with the latter being nested under the machine-paced mode.

Table 1.

Test Design

Test Group	Subjects (White; Nonwhite)	Pacing	Stimulus List
1	44 (31; 13)	Self	I
2	43 (30; 13)	Machine	I
3	40 (31; 9)	Machine	II
4	33 (29; 4)	Machine	III

Procedure

Subjects were brought into the testing room in groups of not more than six. The apparatus was displayed, and the experimental nature of the testing explained briefly prior to issuing the verbal instructions. Instructions emphasized the nature of the stimuli, what was required of the subject in the way of response, and the operation of the apparatus itself. Subjects were then given the opportunity to ask questions about the test and procedure, and to decline participation if they so desired. They were then asked to wait outside the room and were brought in for testing one by one. The instructions for the test were then reviewed with each individual as he was seated at the response box prior to commencement of the experiment.

Stimulus pairs were then presented one by one on the viewing screen for his test condition. Each group of six pairs was presented in 10 consecutive trials with no break between groups. As a subject selected the figure in each pair that he thought was correct, he pressed the corresponding (right or left) response button in front of him. Correct responses were reinforced by a small light in front of the view screen, while incorrect responses received no reinforcement.

As testing was in progress, the experimenter stood behind the subject and recorded his responses on an answer sheet. Responses were also recorded electrically on a two-channel Brush recorder. Upon completion of the test, the subject was cautioned not to discuss anything he had seen or done in the test with those who had not yet been tested. This request was repeated to the entire group after all had been through the test.

Performances by six of the original 160 subjects were discarded. Improper operation of the self-pacing buttons that put the tape reader out of phase with the projector was cause for rejection of three performances. Another subject in the first (self-paced) group was unable to follow instructions. Timer malfunction caused two performances in the first machine-paced group to be eliminated.

Seventeen other subjects' performances were not used in the data analysis because of their Navy Basic Test Battery (BTB) scores and/or demographic data could not be retrieved from computerized records. As a result of these subject losses, the 137 remaining subjects (white and nonwhite) were distributed as follows: Group 1 (24, 11); Group 2 (25, 12); Group 3 (28, 8); and Group 4 (30, 3).

RESULTS

Individual performances in the test, in the form of number of correct choices made per trial per unit of time, were computed to arrive at the test measure of effectiveness, Information Processing Rate (IPR). Specifically, IPR was defined as bits of information correctly processed per second. Performances in the first trial were not used, since responses in the initial trial were dependent wholly upon chance, and as such were not indicative of learning ability.

The number correct in each trial was divided by the amount of time the stimuli were presented to the subject. (In the machine-paced mode, this was a constant 3 seconds per pair. Scores for the self-paced group were scaled to individual rates.) In both situations, the 1-sec. cycle time (inter-stimulus time) of the slide projector was not included in computing IPR. The resultant trial IPR scores were grouped into three blocks of three consecutive trials each. These figures are listed in Table 2. Rates of processing information are seen to generally increase over blocks of trials for all groups. (The single exception is the nonwhite subset of Test Group 4, where performance declines very slightly over trials. This group contained three subjects.) Overall performances by all groups were quite similar, despite differences in pacing mode and stimulus similarity between groups. Overall performance by the nonwhites in Test Group 1 (self-paced) exceeded that of the whites; the reverse was true for the three machine-paced groups. Figures 5 and 6 depict aspects of these situations.

The results listed in Table 2 were subjected to an analysis of variance using a three-way design compensating for unequal cell populations by test group, racial group, and blocks of trials as described by Kirk (1968). The results of this analysis are presented in Table 3. Significant effects were noted between racial groups and among blocks of trials. The blocks effect is important from the construct validity standpoint in demonstrating that learning did occur over all conditions of the experiment. It should also be noted that pacing mode and

Table 2

Information Processing Rate in Multiple Discrimination Learning
by Test Group, Blocks of Trials, and Racial Group

Test Group	Block 1		Block 2		Block 3		Totals		
	White	Nonwhite	White	Nonwhite	White	Nonwhite	White	Nonwhite	Combined
1	181.416	198.000	210.666	227.818	251.958	235.636	214.291	220.484	216.580
2	202.520	156.916	261.640	193.916	284.000	237.333	249.386	196.055	232.089
3	190.071	205.500	240.892	207.750	266.607	212.500	233.713	228.129	232.472
4	187.418	222.000	247.074	215.666	262.814	214.333	231.962	230.533	231.819
<u>Totals</u>									
White	190.009	—	240.509	—	267.461	—	232.659	—	—
Nonwhite	—	187.382	—	210.058	—	228.999	—	208.813	—
Combined	189.361		233.000		257.984				226.783

Note. Entries are bits/sec X 10^3.

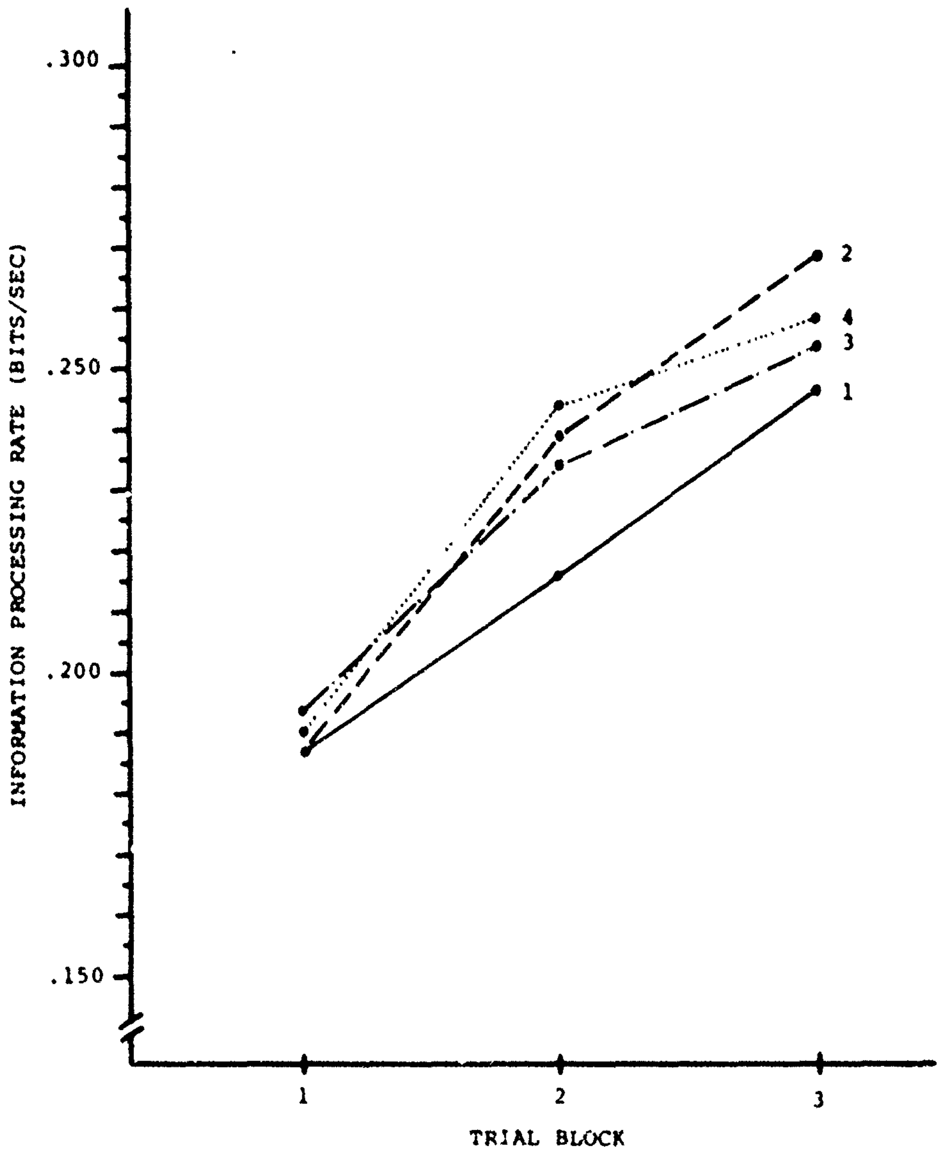


FIGURE 5. Information Processing Rate by Test Group and Blocks of Trials.

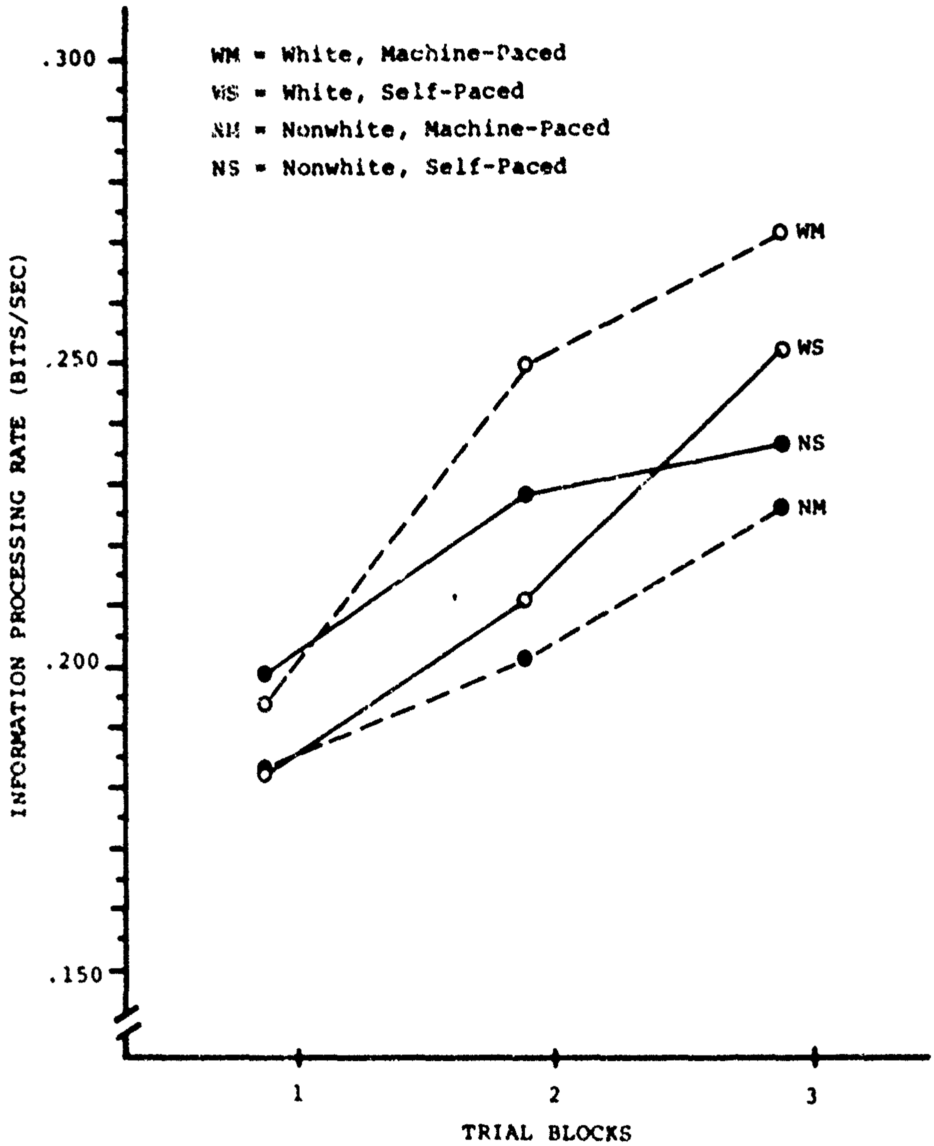


FIGURE 6. Information Processing Rate by Racial Group, Pacing Mode, and Blocks of Trials.

Table 3.

Analysis of Variance of Overall Performance by Test Group, Racial Group, and Blocks of Trials

Term	df	SS	MS	F	p
Total	243	1,829,659.50	—	—	—
Test Group (T)	3	5,082.50	1,694.10	0.230	n.s.
Racial Group (R)	1	31,511.00	31,511.00	4.288	<.05
Trial Block (B)	2	117,910.00	58,955.00	8.023	<.001
T X R	3	24,396.00	8,131.90	1.106	n.s.
T X B	6	10,165.00	1,694.10	0.230	n.s.
R X B	2	21,346.00	10,673.00	1.452	n.s.
T X R X B	6	13,214.00	2,202.30	0.299	n.s.
Error	220	1,616,200.00	7,347.60	—	—

similarity were confounded in the test group variable in this analysis, but had the primary effects of either of these variables been substantial, the analysis would have resulted in a significant F for the test group variable. On the other hand, if the effects of both variables had been substantial, the effects on the test group variable would have been indeterminate because of the possibility that the effects of one might cancel the effects of the other.

In order to assess the effects of pacing mode, an analysis of variance was conducted using the total IPR as the dependent measure and racial group and pacing as the independent variables. Racial group was included in the analysis because of the possible interactive effect with the pacing variable, as suggested in Figure 6. With the data collapsed over blocks of trials, the racial variable was not significant (Table 4). The pacing effect was not significant and the hypothesized interactive effect attained a F value that was between the .10 and .20 levels of probability.

In order to assess a possible similarity effect, an analysis of variance was conducted using the total IPR as the dependent measure and racial group and similarity (stimulus set) as the independent variables. Only the machine-paced test groups were used for this analysis. The results, shown in Table 5, found racial group to be significant at less than the 2 percent level of probability, while similarity and the interaction term were not statistically significant. In addition to the implications for the similarity variable, the comparative analysis provided by tables 4 and 5 with respect to race indicate that race did have a significant effect when the subjects were machine-paced but not when they were allowed to pace themselves.

Finally, in order to confirm that subjects showed a significant difference in their learning rates, as one would expect from the sizable error terms in all of the preceding analyses, several analysis of variance tests were conducted using a repeated measures design with subjects and blocks of trials as the independent variables and the interaction of these two effects as the error term. The dependent variable was the IPR per subject per block. Four such tests were conducted by partitioning the total sample by race and pacing mode. The F ratios were all highly significant for subjects and blocks of trials with most of them at the .001 level of probability.

Internal reliability of the test itself was investigated using a split-half design for each test group and each racial group as well as for overall performances. Processing rates were compared for trials 4, 6, and 8 against those of trials 5, 7, and 9. In addition, scores on the latter group of trials were compared with those obtained on trials 6, 8, and 10. The former comparison will be referred to as "low trials" and the latter, as "high trials."

Correlation coefficients thus obtained were used in the Spearman-Brown formula for split-half correlations. Both the raw coefficients

Table 4

Analysis of Variance of Overall Performance by Racial Group and Pacing Method

Term	df	SS	MS	F	p
Total	137	465,094.994	—	—	—
Racial Grp (R)	1	4,417.475	4,417.475	1.310	n.s.
Pacing Mode (P)	1	242.501	242.501	0.071	n.s.
R X P	1	8,772.961	8,772.961	2.602	n.s.
Error	134	451,662.057	3,370.612	—	—

Table 5

Analysis of Variance of Overall Performance by Racial Group and Stimulus Set (Machine - Paced Only)

Term	df	SS	MS	F	p
Total	102	5,316.928	—	—	—
Racial Grp (R)	1	342.169	342.169	6.810	.020
Stimulus Set (S)	2	4.758	2.379	0.047	n.s.
R X S	2	96.117	48.058	0.956	n.s.
Error	97	4,873.884	50.246	—	—

Table 6

Split-Half Reliability Coefficients

Group		Low Trials (468 vs 579)		High Trials (579 vs 6810)		Totals	
		r(raw)	r(S-B)	r(raw)	r(S-B)	Low	High
1	White	.767	.868**	.713	.832**	.865**	.872**
	Nonwhite	.756	.861**	.864	.927**		
2	White	.800	.889**	.865	.928**	.871**	.921**
	Nonwhite	.700	.824**	.826	.905**		
3	White	.615	.762**	.632	.775**	.722**	.759**
	Nonwhite	.367	.537	.535	.697*		
4	White	.674	.805**	.664	.798**	.802**	.794**
	Nonwhite	.637	.778	.610	.758		
Totals							
	White		.835**		.843**		
	Nonwhite		.788**		.873**		
	Combined		.824**		.851**	.838**	

*Significant at $p < .05$.

**Significant at $p < .01$.

and the Spearman-Brown coefficients are listed in Table 6. A majority of the coefficients are seen to be statistically significant.

The relationship between scores on the experimental test and the traditional methods of measuring Navy recruit potential was investigated using the test subjects' scores on the Navy General Classification Test (GCT), a major portion of the standard Basic Test Battery (BTB). The basis for the GCT lies in verbal ability, since the test consists of sentence completions and verbal analogies. Test scores are scaled on a normalized distribution with a mean of 50 and a standard deviation of 10. Performance on the Arithmetic Reasoning Test (ARI) is often combined with GCT scores to obtain a rough "multiple" used in determining Navy technical school eligibility and aptitude.

Pearson product-moment correlations were computed between test scores and GCT scores obtained from individual service files. (One nonwhite subject was dropped from this analysis because his GCT score was not available.) These correlations were determined for racial subgroups of subjects falling below and above the GCT mean score of 50, for both racial groups *in toto*, and for the entire sample. These figures are seen in Table 7. Significant values of the correlation coefficient are noted only in the white group as a whole and for the entire sample. Nonwhite test scores did not correlate significantly with GCT performance.

Table 7

Correlations of Test Performance (IPR) with Navy General Classification Test (GCT) Score

Group		Group Averages		Correlation Coefficient		
		GCT	IPR	GCT GRP	Race GRP	Total
Nonwhite $N=33$	Low (<50) $N=24$	42.67	.208	.316	.213	.270**
	High (≥50) $N=9$	56.89	.207	.601		
White $N=104$	Low (<50) $N=17$	42.18	.207	.253	.223*	
	High (≥50) $N=87$	59.63	.238	.050		

*Significant at $p < .05$.

**Significant at $p < .01$.

DISCUSSION

CONSTRUCT VALIDITY

The test was constructed to be a measure of learning ability with the implication that learning ability is a manifestation of the intellectual capacity of a person. Differences in this intellectual capacity between individuals was assumed to be measurable by the rate with which new material is learned. Using IPR as the rate measure, the results of the trial administration of the test showed that learning took place and that the rate was different among individuals. Moreover, the results were found to be highly reliable--especially for a 4-minute test--using an internal (split-half) criterion of reliability. Thus, the basic essential requirements for the construct validity of the test would seem to have been adequately demonstrated. Additional experimentation would be required to show that it is, indeed, a differential measure of intellectual capacity. Probably the best way to demonstrate this essential requirement would be to give the test to different age groups. The fact that the items had been standardized for information content (1 bit per item) would make it possible to administer shorter forms of the test--e.g., four instead of six items--to different age groups and yet have the IPR mean the same when corrected for total information content of the stimulus lists.

Earlier in this paper, it was stated that the construct validity of a test required an answer to the question, What does the test measure? The answer given here is learning ability. But, as Estes (1974) has argued, a product-defined measure of intelligence or ability does not provide an understanding of what intelligence is. Rather, the process should be defined and the relationship between the process and the product measure should be determined. The design of this trial administration of the test does not provide opportunities to answer the process question. Since similarity, however, was not a significant variable, visual discrimination of the stimuli would not seem to have been involved in the learning process. Based on a great deal of research in recent years in the area of human learning and information processing, it would be safe to say that some form of coding of the individual forms and, probably, the stimulus pairs as an entity was required. Additionally, short-term memory was required to hold the information pertaining to one item in working memory while processing a new item. Here, some sort of mnemonic device might be involved, and in both cases verbal fluency and image formation might be the basic skills underlying these processes. With respect to verbal ability playing a role, the small, significant correlation between IPR scores and the GCT scores for the white group would support this contention. Taken in conjunction with this finding, the absence of a significant correlation for the nonwhite group could also be seen as not disconfirming the trend, if it is assumed that the GCT score is not as good a measure of verbal ability for subjects in the nonwhite group. These results, however, only emphasize that the measure of verbal fluency or the capacity to generate useful images must be appropriate to the cultural background of the individual subject.

CULTURAL IMPLICATIONS

If the subjects--white and nonwhite--had comparable learning abilities, no racial group differences would be found on the IPR. The study found no significant differences among the self-paced subjects, but a significant difference was found for racial groups in the machine-paced mode. A problem in attempting to determine from the experiment data whether the white and nonwhite group differed in learning ability lies in the fact that the subjects were a selected group that was not representative of America's youth in general. As noted, the average education level was at the 12th grade. The information in Table 7 shows that 60 percent of the sample was above the median in GCT scores. There was a considerable difference in racial groups, however, with 84 percent of the white group being above the 50th percentile, whereas only 27 percent of the nonwhite subjects were in that category. There was a small but significant correlation of GCT scores with the IPR, but only for the white group and the entire sample. How can these data be related to the cultural implications of the test?

With respect to the differences noted in the paced and self-paced groups, it may be that the machine-paced format placed greater pressure on the subjects and generated greater test anxiety. Where short-term memory and the learning of discriminations involving very similar items constitute the task, the effects of anxiety could be disruptive as shown by Taylor and Spence (1952) and Ramond (1953) in serial, verbal learning tasks. For anxiety to have a differential effect in the racial groups, the anxiety induced by the test conditions would have to be greater for the nonwhite group. This could be true as a part of the larger picture of differences in test-taking motivation, attitudes, experience, and skill that have been attributed to different cultural backgrounds If these contentions are valid, then the self-paced mode would be more culture-free in its assessment of the test subject. If the finding in this trial administration of the test for the self-paced condition should hold up in subsequent administrations, then this would be strong evidence for the culture-fair nature of this test.

The pattern of correlations between the IPR and the subjects' GCT scores takes the form that Jensen (1968b) found with children of high and low socioeconomic (SES) groups. Noting that children from low SES backgrounds with IQs in the range of 60 to 80 appear to be much brighter in social and nonscholastic behavior than their middle- or upper-middle SES counterparts, he gave groups of such children learning tasks in the laboratory and compared their learning performance with standard intelligence test scores for the children. There was a substantial correlation of IQ and learning scores for middle-class children, but the correlation was negligible for children from low SES backgrounds. Jensen attributed the difference to the fact that the learning tasks and the intelligence tests measured two different levels of intelligence with the lower level, measured by the learning tasks, being common to both groups and the other being better represented within the high SES group. In the present instance, it would seem more parsimonious to conjecture that the IPR was a measure of intellectual capability for both

groups, whereas the GCT, which has been found to be culturally biased (Stephan, 1973; Thomas, 1972c), was a fair measure only for the white group. In addition, the significant correlations accounted for only a very small portion of the variance in IPR scores. Accordingly, it would appear that the multiple discrimination test is indeed culture fair and provides an unbiased measure of learning ability, at least in the self-paced form. Larger and more numerically balanced samples from an unselected population would be necessary to confirm these conclusions.

TEST AND TESTING CONSIDERATIONS

Discussion in this section will deal with the psychometric and physical aspects of the multiple discrimination learning test. Specifically, the length of the test, additional matters pertaining to the pacing mode, and the physical packaging of the test will be considered.

Test Length

The decision to stop the test after 10 trials was arbitrary. Several subjects showed errorless performance within this limitation. In the machine-paced mode where there was a theoretical limit to the IPR of .333 bits/sec., examination of the third block of trials showed that the white subjects attained a maximum of 80 percent of this perfect learning rate, while nonwhites reached 69 percent of this quantity. While it is not possible to tell how many trials are required for perfect learning, since a trials-to-criterion design was not used, it would be advisable from a psychometric standpoint to stop short of perfect learning when the difference in learning rate among subjects is more variable. There would also be a tradeoff between a test length of maximum discriminability among subjects and one of highest reliability, which might not be the same. Thus, the optimum test length is not a simple question that yet remains to be determined.

Pacing Mode

It has been previously shown that pacing mode appeared to have a difference on test results with the self-paced mode being more culture-fair. From a psychometric standpoint, the difference between the two methods is that the self-paced mode places no limit on the IPR that a subject might attain. This would lead to greater variability among subjects and, presumably, a more reliable differentiation among test takers. Since many more variables are free to exert their effects with the self-paced mode, it may be, however, that less reliable performance may result. The self-paced mode, though, should be more representative of the manner in which a subject approaches and deals with a problem, and the results of the testing, as a consequence, would be more generalizable to real-life situations where learning is required. That is, it should permit greater predictive validity.

The self-pacing mode would also be desirable on the basis of the discussion on the construct validity of the test. There it was stated that the rate of learning would be the measure of learning ability, and the self-paced mode is the only one that permits an assessment of this measure. The highest rate in this study was .503 bits/sec., which occurred in the nonwhite subgroup of the self-paced condition. Accordingly, the self-paced mode would appear to be the better procedure for this test.

Physical Packaging of the Test

The type of stimulus materials, their presentation method, and scoring make it relatively simple to institutionalize the test using teaching machines with true-false or multiple-choice response provisions. Scoring counters could be readily integrated with the machine. With the ever-expanding use of computer terminals at remote locations, the test could easily be set up to be administered from a central location. This would permit the ready selection of a test "form" from among several that could be accessed, and scoring and performance analysis would be almost instantaneously provided upon completion of testing.

A specific item that requires improvement over the set-up used in this trial administration of the test is the advance procedure in the self-paced mode. In this trial, the subject had to call for the next stimulus after responding by pressing a button on the side of the response unit. As a result, learning times for the self-paced group might have been slightly biased upwards.

Another feature that requires investigation is whether the reinforcement should be given by a signal only for correct choices. That was the procedure in this trial administration. The learning literature has a large number of studies that have investigated positive reinforcement, negative reinforcement, both positive and negative reinforcement, and correction vs. noncorrection methods--e.g. Arima (1965). There is a good likelihood that the correction method might be best for this test. That is, the next stimulus item will not appear until the subject presses the correct button. If the subject has initially chosen the incorrect alternative, he or she must press the correct button. The best mode should be determined by experimentation.

SUMMARY AND CONCLUSION

The purpose of this study was to develop a test of learning ability that would not be affected by the cultural background of the individual being tested. A test was created using randomly shaped, 2-dimensional polygons presented in pairs in a discrimination learning paradigm. Three different lists of six such pairs were created so that multiple discrimination learning was involved. The lists were presented individually in a manner similar to verbal discrimination learning in both a self-paced and machine-paced mode.

In a trial administration of the test using Navy recruits as subjects, significant learning took place over 10 trials. Nonwhite and white racial groups, which differed significantly on their Navy General Classification Test Scores, performed at a comparable level in the self-paced mode. The adjusted reliability of the test (split-half) was .85. The correlation of the test scores with the GCT scores was marginally significant for the white group and the total sample, but not for the nonwhite group. There was no difference in performance among the three lists, which differed considerably in the similarity of the stimulus materials. This suggested that any combination of the forms could be used to create equivalent alternate forms.

It was concluded that a practical test of learning ability that was culture fair to both the white and nonwhite groups had been demonstrated. Refinement of the test would be desirable with respect to optimal length, reinforcement procedure (correction vs. noncorrection), and the physical packaging of the test.

REFERENCES

Arima, J. K. Human probability learning with forced training trials and certain and uncertain outcome choice trials. Journal of Experimental Psychology, 1965, 70, 43-50.

Arima, J. K. Verbal discrimination learning: An analysis of randomly presented 2-, 3-, and 4-word items. JSAS Catalogue of Selected Documents in Psychology, 1974, 4, 116.

Arnoult, M. D. Familiarity and recognition of nonsense shapes. Journal of Experimental Psychology, 1956, 51(4), 269-276.

Bitterman, M. E. The evolution of intelligence. Scientific American, January 1965, 92-100.

Bitterman, M. E. The comparative analysis of learning. Science, 1975, 188, 699-710.

Congressional Budget Office. The costs of defense manpower: issues for 1977. Washington: Superintendent of Documents, January 1977.

Eckert, E. and Kanak, N. J. Verbal discrimination learning: a review of the acquisition, transfer, and retention literature through 1972. Psychological Bulletin, 1974, 81, 582-607.

Estes, W. K. Learning theory and intelligence. American Psychologist, 1974, 29, 740-749.

Feldman, M. W., and Lewontin, R. C. The heritability hang-up. Science, 1975, 190, 1163-1168.

Goslin, D. A. Standardized ability tests and testing. Science, 1968, 159, 851-855.

Green, D. R. What does it mean to say a test is biased? Education and Urban Society, 1975, 8, 33-52.

Green, E. J., and O'Connell, J. A. An annotated bibliography of visual discrimination learning. Teachers College Press, Columbia University, New York, N. Y., 1969.

Holtzman, W. H. The changing world of mental measurement and its soci l significance. American Psychologist, 1971, 26, 546-553.

Hunter, J. E., Schmidt, F. L., and Rauschenberger, J. M. Fairness of psychological tests: implications of four definitions for selection, utility and minority hiring. Journal of Applied Psychology, 1977, 62, 245-260.

Jensen, A. R. Social class, race, and genetics: Implications for education. American Educational Research Journal, 1968, 5, 1-42. (a)

Jensen, A. R. Patterns of mental ability and socioeconomic status. Proceedings of the National Academy of Sciences, 1968, 60, 1330-1337. (Reprinted in American Psychologist, 1968, 26, 1330-1337.) (b)

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Calif.: Wadsworth, 1968.

Matarazzo, J. D. Wechsler's measurement and appraisal of adult intelligence. Baltimore: Williams and Wilkins, 1972.

Melton, A. W. Military psychology in the United States of America. American Psychologist, 1957, 12, 740-746.

Ramond, C. K. Anxiety and task as determinants of verbal performance. Journal of Experimental Psychology, 1953, 46, 120-124.

Stephan, R. A. Evidence of racial bias in military testing. Paper presented at 31st meeting of Military Operations Research Society, Monterey, California, 1973.

Taylor, J. A., and Spence, K. W. The relationship of anxiety level to performance in serial learning. Journal of Experimental Psychology, 1952, 44, 61-64.

Thomas, P. J. The relationship between Navy classification test scores and final school grade in 104 Class "A" schools (Res. Rep. SRR 72-15). San Diego: U.S. Navy Personnel and Training Research Laboratory, 1972. (a)

Thomas, P. J. The relationship between Navy classification test scores and final school grades in 98 Class "A" schools (Res. Rep. SRR 72-22). San Diego: U.S. Navy Personnel and Training Research Laboratory, 1972. (b)

Thomas, P. J. An investigation of possible test bias in the Navy Basic Test battery (Tech. Bull. STB 73-1). San Diego: U.S. Navy Personnel and Training Research Laboratory, 1972. (c)

Windle, C. and Vallance, T. R. The future of military psychology: paramilitary psycholocy. American Psychologist, 1964, 19, 119-129.

Young, P. A. A culture-free performance test of general learning ability. M.S. Thesis, Naval Postgraduate School, Monterey, Calif., December 1975.

Cutting Scores--Legal Implications

Lawrence S. Buck

Abstract

A vital step in the process of test development is determining whether the test in question requires that a standard be set, i.e., that a cutting score is needed. Cutting scores are generally considered essential for criterion-referenced tests (CRTs) while for norm-referenced tests (NRTs) the issue is less clear. Cutting scores cannot always be strongly justified on psychometric grounds for NRTs but are often justified on other grounds such as legal issues, administrative reasons and probabilistic terms. Therefore, both CRTs and NRTs will usually require the setting of a cutting score.

The process by which the cutting score is determined as well as the particular choice of cutting score will have legal implications for the test as a whole. As employee selection practices and procedures come under increasing scrutiny and challenge in the courts, issues related to the determination and setting of cutting scores will also be scrutinized.

This report looks at some of the legal implications of setting cutting scores for both CRTs and NRTs in terms of professional guidelines, accepted practices and related court decisions. Because legal precedence is one of the basic foundations of American law, a close look will be taken at some court cases that have at least in part been concerned with cutting scores. This is done with an eye towards determining if or what legal precedence exists and what positions the various courts that have dealt with cutting score issues have taken.

Cutting Scores--Legal Implications

A vital step in the test development process is determining whether the test in question requires that a standard be set, that is, that a cutting score is needed. The issue is not always clear as to whether cutting scores are in fact necessary. While cutting scores are generally considered to be essential for criterion-referenced tests (CRTs), for norm-referenced tests (NRTs) the issue is less clear. Cutting scores cannot always be strongly justified on psychometric grounds for NRTs. With NRTs, it is often difficult to determine from the number of test questions answered correctly which score should distinguish qualified applicants from unqualified applicants. However, the use of cutting scores with NRTs is often justified on other grounds such as legal issues, administrative reasons and/or probabilistic terms. In situations in which minimum standards are thoroughly and validly determined to be essential to success on the job, as is often the case with CRTs, little issue can be taken with the cutting scores selected as long as they accurately reflect the minimum standards. Therefore, one usually finds cutting scores used with both CRTs and NRTs.

The term "cutting score" is often assigned a variety of meanings, often depending on the test and situation in which a score(s) is designated to be a cutting score(s). For purposes of this report, "cutting score" refers to an established standard and is considered to be synonymous with other terms such as: qualifying score, critical score, passing point, cutoff score and cut score when these terms also refer to an established standard. Regardless of the way in which one chooses to define cutting scores or the methods used to set cutting scores, it is important to point out that the process of selecting or setting a cutting score(s) is not totally objective for either NRTs or CRTs. As Ebel (1972) so aptly points out,

> Anyone who expects to discover the 'real' passing score... is doomed to disappointment, for a 'real' passing score does not exist to be discovered. All any examining authority that must set passing scores can hope for, and all any of their examinees can ask, is that the basis for defining the passing score be defined clearly, and that the definition be as rational as possible. (p. 496)

While there will always be a need for some human judgment in the process of setting a cutting score there are many aspects of the process that can be realistically and appropriately determined. Any cutting score(s) used must be thoroughly justified and documented.

It is essential that any cutting score(s) used must be defensible as employee selection procedures and practices have come under increasing scrutiny in recent years, especially in regard to such issues as test bias, adverse impact and discrimination. Cutting scores (when used) are a vital, although often overlooked, part of the selection process and as such are also subject to careful examination. Many employee selection practices have been and are presently being challenged through the courts by minority groups. Where the selection process is at least in part based on normative standards, the choice of any cutting score(s), no matter how it is determined, may be susceptible to attack. It is the responsibility of the test developer to ensure that any cutting score(s) is properly documented and supported.

The crucial question then becomes, how does the test developer ensure that a cutting score(s) is properly and adequately supported? Where can test developers look for guidance in this area and what kinds of guidance can they expect to find? What are the legal implications of setting cutting scores for CRTs and/or NRTs in terms of professional guidelines, accepted practices and related court decisions? Increasing interaction with the legal system of our country is having and will continue to have a great impact on the field of testing. Legal decisions have in the past and will in the future play a role in the manner in which the subject of cutting scores is treated. A basic foundation of American law is the concept of precedence and it is to the courts that test developers will look for legal guidance in the use of cutting scores. In addition, the test developer must also look to professional guidelines for advice and guidance on cutting score issues.

Before delving into some of the legal decisions that impact on cutting score issues, let us first look at cutting scores as they are dealt with in some professional guidelines and regulations. The "Federal Executive Agency Guidelines on Employee Selection Procedures" state that:

> where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force. If other factors are used in determining cutoff scores, such as the relationship between the number of vacancies and the number of applicants, the degree of adverse impact should be considered. (Federal Executive Agency Guidelines on Employee Selection Procedures, published simultaneously by the U.S. Civil Service Commission, the U.S. Department of Justice and the U.S. Department of Labor, 1976, p. 51753)

The Equal Employment Opportunity Commission's "Guidelines on Employee Selection Procedures" (1976) treat the subject of cutting scores as follows:

> ...for each test that is to be established or continued as an operational employee selection instrument, as a result of the validation study, the minimum acceptable cutoff (passing) score on the test must be reported. It is expected that each operational cutoff score will be reasonable and consistent with normal expectations of proficiency within the work force or group on which the study was conducted. (p. 51985)

In addition,

> ...where a test is valid for two groups but one group characteristically obtains higher test scores than the other without a corresponding difference in job performance, cutoff scores must be set so as to predict the same probability of job success in both groups. (p. 51585)

And, in discussing continued use of tests which are not fully supported by the required evidence of validity:

> It is expected also that the person may have to alter or suspend the cutoff scores so that score ranges broad enough to permit the identification of criterion-related validity will be obtained. (p. 51986)

The "Principles for the Validation and Use of Personnel Selection Procedures" published by the Division of Industrial-Organizational Psychology American Psychological Association (1975) has the following to say about cutting scores:

> If cutting scores are used as a basis for decision (i.e., as rigid pass-fail points) the rationale or justification should be known to all users. This principle does not recommend cutting scores. Rather, 'The intent is to recommend that test users avoid the practice of designating purely arbitrary cutting scores they can neither explain nor defend.' If cutting scores are to be established, some consideration should be given to the different effects

> of different cutting scores; e.g., the effects of the two kinds of error: selecting people who prove unsatisfactory as opposed to rejecting people who would have been satisfactory if hired. (p. 14)

The "Standards for Educational & Psychological Tests" (1974) discuss cutting scores as follows:

> If specific cutting scores are to be used as a basis for decisions, a test user should have a rationale, justification, or explanation of the cutting scores adopted. ...The test user should have some justifiable reason for the adoption of a given cutting score. ...This standard does not attempt to recommend a specific procedure for developing cutting scores where they are to be used. The intent is to recommend that test users avoid the practice of designating purely arbitrary cutting scores they can neither explain nor defend. (pp. 66-67)

The Standards suggest a variety of ways of selecting a cutting score. For example, with content-referenced interpretations of mastery tests, the cutting score "might be determined as the obtained score at which one can reject, at a preselected level of probability, the hypothesis that a predesignated confidence interval for that score includes the perfect score on the test" (p. 66). In other situations the cutting score(s) might be based on "a designated probability of achieving a specified level of success;" the score "that will maximize the discrimination between high- and low-criterion groups;" or "on a distribution of scores in a 'predicted-yield' situation" (pp. 66-67).

The guidelines and regulations that are quoted above do not give a great deal of coverage to the cutting score issues. However, to summarize the main points, they generally stress that cutting scores: should never be purely arbitrary; should be reasonable and consistent with normal expectations of acceptable proficiency within the work force; and, they should be justified and documented.

As I previously pointed out, the concept of precedence is a basic foundation of American law. As legal challenges increase, test developers must look to the courts for legal guidance on cutting score issues. A number of court cases are reviewed below that have, at least in part, dealt with cutting score issues. This is done with an eye towards determining if the various court decisions have delineated any clear and consistent guidelines for cutting scores. It will be readily apparent that there is not

a great deal of agreement among the various court decisions in the stances that they have taken towards cutting score issues. In fact, many court decisions seem to contradict each other. In many court cases that are relevant, cutting scores have been discussed only in very general terms. It is also true, and unfortunately so, that the judicial lawmaking process tends to proceed on a case-by-case basis with little or no input from any situations which have not become involved in a court case. No one can be sure what the courts will do in the face of new arguments. At this point in time, we are unprepared to elucidate a selection theory, in terms of cutting score issues, from the legal decisions that have to date been handed down.

At the present time one of the more important legal decisions in the area of testing is the U.S. Supreme Court decision in the Griggs v. Duke Power Company case. Footnote 11 of this case states in part that:

> ...an employer may set his qualifications as high as he likes, he may test to determine which applicants have these qualifications, and he may hire, assign and promote on the basis of test performance. (p. 6435)

In some cases the courts have held cutting scores to be legal if the users present evidence that the appropriate subject matter experts endorse them. For example, in Tyler v. Vikery, a case in which the passing score for a bar examination was challenged the cutting score was evaluated as follows:

> ...There (in Armstead v. Starkville Municipal Separate School District, 461F 2d 276 5th. Cir. 1972) we suggested that a rationally supportable examination should 1) be designed for the purpose for which it is being used, and 2) utilize a cutoff score related to the quality the examination purports to measure. Both the essay and the MBE portions of the examination are designed solely to assess the legal competence of bar examinees and while the minimum passing score of 70 has no significance standing alone, it represents the examiners considered judgments as to 'minimal competence required to practice law,' the precise quality the exam attempts to measure.

A district judge in Los Angeles ruling on height and physical agility cutoffs for the Los Angeles police department ruled that such requirements do not violate either the Constitution or Title VII of the Civil Rights Act. In referring to evidence in the form of criterion-related validation

studies the court stated:

> ...A stronger and taller officer can more quickly and effectively control another person in the event of a confrontation. This officer is more likely to be able to do so without having to resort to extreme force, such as the use of a gun or a dangerous control hold. ...As defendants' affidavits clearly show and common sense confirms, physical size, strength, and agility have a direct relation to the quality of the performance of these functions. ...These qualities are essential to an effective, efficient, and functional police force.

In addition the court stated:

> Since the tests are job related, the setting of cut-off scores is a matter for the employer's judgment. Many factors may go into this decision, such as the number of applicants available, the cost of failures in training and on the job, the critical nature of the job to be performed, and the level of performance at which the employer desires employees to perform.

While the above cases were generally in favor of the use of cutting scores, the following court cases have to one degree or another ruled against the use of particular cutting scores. For example, in Rogers v. International Paper Company (1975), cutting scores were judged to be too high in that 40 percent of the skilled craftsmen in the sample would not have been able to achieve admission to their respective crafts under this standard. In Hiatt v. Berkeley (1975), the California Superior Court ruled that the method in question of using tests as a pass/fail standard was unjustified and arbitrary; tests that are not found to be job-related should be evaluated on the basis of achievement. In U.S. v. Central Motor Lines, Inc. (1971), it was ruled that the determination of the passing score was at the subjective discretion of the employer on an individual by individual basis and thus the test was unlawful.

The use of the National Teacher Examination as a licensing examination for public school teachers in North Carolina was declared unconstitutional in U.S. v. State of North Carolina (1975). The basis for the decision was that the establishment of the cutting score was considered to be arbitrary and was not shown to be a measure of the minimum standard for the teaching

profession. While the court did not find anything wrong with the test itself or the use of a cutoff score per se, it did state that:

> such cutoff score shall first have been validated with respect to minimum academic knowledge an applicant must possess in order to become a reasonably adequate and competent teacher and that such score be shown to bear a rational relationship to teaching capacity. (Employment Practices Decisions, 1975, p. 1)

Apparently the court felt that the state had selected a score calculated to produce a given failure percentage for its cutting score, a practice which was in this case declared unlawful.

The cases cited in this paper are neither a comprehensive sampling of all cases in which cutting scores are at least an issue, nor are they necessarily representative of all such cases. However, they do represent some important legal cases in which cutting scores have been an issue and as such they warrant some consideration. In addition, they rather vividly illustrate the discrepant decisions and philosophies that are being taken in regard to cutting score issues. It would be necessary for one to look at the total picture for each particular case before claims for similarities or dissimilarities between the various court decisions could be made. However, from a review of the cited cases and others not here included, it seems evident that some basic philosophical differences do exist among the various courts.

One crucial issue that does tend to stand out amidst the confusion is the requirement stated in many cases that test users must demonstrate the validity and job-relatedness of their tests. For the most part, while the various courts have not ruled out the use of cutting scores per se, they have insisted that such scores be properly documented and supported in terms of job relatedness.

It is hoped that professional test developers will in the future have more of an opportunity to provide some input on cutting score issues to the courts. This may lead to more of a concensus concerning these issues and better guidelines and regulations to follow when setting cutting scores.

References

American Psychological Association, Inc. Standards for educational and psychological tests (Rev. ed.). Washington, D.C.: Author, 1974.

Division of Industrial-Organizational Psychology, American Psychological Association. Principles for the validation and use of personnel selection procedures. Dayton, Ohio: The Industrial-Organizational Psychologist, 1975.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

Equal Employment Opportunity Commission. Guidelines on employee selection procedures. Federal Register, 41(228), 1975.

Federal executive agency guidelines on employee selection procedures. Federal Register, 41, 1976, 51733-51759.

Griggs v. Duke Power Company, Footnote 11, 91 Supreme Court 849.

Psychological Services Incorporated Newsletter, 1(2), 1977. Los Angeles District Court Memorandum.

Rogers v. International Paper Company. Employment Practices Decisions, 9, 1975.

Tyler v. Vikery. Employment Practices Decisions, 10, 1975.

U.S. District Court, North Carolina. U.S. v. State of North Carolina. Employment Practices Decisions, 10(438), 1975.

U.S. v. Central Motor Lines, Inc. Employment Practices Decisions, 4, 1971.

AN EVALUATION OF SELECT APPROACHES
FOR BIASED ITEM IDENTIFICATION

Lawrence M. Rudner
Gallaudet College, Model Secondary School for the Deaf

and

John J. Convey
Catholic University of America

Paper presented at the annual meeting of the Military Testing Association, San Antonio, Texas, October, 1977

Problem

Approximately 25 years ago, Eels and his colleagues conducted what appears to be the first serious attempt to examine test items for bias (Eells, Davis, Havighurst, Herrick and Tylar, 1951) and developed one of the first measures purported to be culture fair. Since that time, the entire issue of cultural bias in measurement has become heated, complex, and pronounced in the literature. Actions by the National Association of Black Psychologists, the American Personnel and Guidance Association of Black Psychologists, the American Personnel and Guidance Association, the National Education Association, the National Assocation for the Advancement of Colored People, the National Association of Elementary School Principlas and the Council of the Society for the Psychological Study of Social Issues calling for moritoria on certain types of tests, banning tests, and requiring alternative plans for testing, indicate the serious nature of the current situation (see Williams, Mosby and Hinsen, 1977). The concern is also apparent in recent litigation (DeFunis vs. Odegaard, 1974; Diana vs. the California State Board of Education, 1970; Hobsen vs. Hansen, 1967). Naturally, all this has not gone unnoticed by those involved in the measurement field. Bias and debiasing studies have occurred and various models been proposed in ever-expanding efforts to meet the challenge of bias in educational assessment.

One major type of bias investigation is concerned with the instrument as a whole and examines the question: Does a test unduly favor or impede examinees from different parts of the country or of different backgrounds? Another is concerned with the items within a test and asks: Which items and item formats are appropriate for a given population and which may be used across given cultures?

The first type of investigation is of interest to the test users who need to evaluate the appropriateness of the test information. The models

proposed by Cleary (1968), Thorndike (1971), Darlington (1971), Cole (1973), Einhorn and Bass (1971) and Gross and Su (1975) (also see the entire Spring 1976 issue of the Journal of Educational Measurement) exemplify this first type of investigation. The second type of investigation is of interest to developers as it assists them in developing valid and cross-culture fair items and provides a framework for constructing better tests in subsequent efforts. By identifying and removing such items from an initial item pool, test developers could, theoretically, develop a measure free of bias. The work of Angoff (1972), Cardall and Coffman (1964), Green and Draper (1972), Merz (1973, 1976), Rudner (1977a), Scheuneman (1975) and Veale and Foreman (1975, 1976) (see the reviews by Merz, 1977 and Rudner, 1977b) have been directed at this need. It is this second type of bias--item bias--which the present paper addresses.

Typically, these researchers have adopted a single approach and used that approach exclusively in their work. As a result, studies applying more than one approach to a single set of data have been sparce. This situation has led to the problem identified by Merz (1977) and addressed by this study: the psychometric properties of the approaches have not been fully evaluated using hypothetical and actual item response data.

Purpose

The purpose of this study was to investigate the following four approaches to biased item identification using common sets of actual item response data:

1. Transformed item difficulties in which within group p-values are standardized and compared between groups (Angoff, 1972);
2. Chi-square in which individual items are investigated in terms of between group score level differences in expected and observed proportions of correct responses (Scheuneman, 1975);

3. Item characteristic curve theory in which differences in the probabilities of a correct response given examinees of the same underlying ability and in different culture groups are evaluated (Rudner, 1977a);

4. Factor score in which item bias is investigated in terms of loadings on biased test factors (Merz, 1973).

The investigation addresses the following questions:

1. Do the select approaches provide identical classifications of items as to their degree of aberrance when applied to item response data corresponding to two culturally different populations?

This question calls for a comparison of the approaches as they would typically be applied in test development or test evaluation studies.

2. Do the select approaches provide classifications of minimal bias when applied to subsamples of a single population?

This question is similar to one asked by Jensen (1973) and serves to evaluate the adequacy of the various approaches. Here, an approach identifying an abundance of items as biased would be suspect as being inadequate.

The Models

Transformed Item Difficulties

This approach, which examines the interaction of item and groups, appears to be one of the best known. It has been advocated and used frequently by Angoff (1972; and Ford, 1973; and Modu, 1973) and others (Green and Draper, 1972; Jensen, 1973; Hicks, Donlon, and Wallmark, 1976; Strassberg-Rosenberg and Donlon, 1975; Echternacht, 1975; Rudner, 1977c).

In this method, p-values for a group of items are obtained for two different groups of examinees. Each p-value is converted to a normal deviate and the apris of normal deviates, one pair for each item, are plotted on a

bivariate graph, each pair represented by a point on the graph.

The plot will generally be in the form of an ellipse. A 45 degree line, passing through the origin, provides an indication of the absence of bias. Items greatly deviating from this line may be regarded as exhibiting an item by group interaction. That is, relative to the other items, deviant items are especially more difficult for members of one group than the other. Assuming both groups received similar instructions, such items would appear to represent different psychological meanings for the two groups of examinees.

Since the intent is to make comparisons of between-group differences in item difficulty, it is necessary to transform the proportion passing an item to an index of item difficulty which constitutes at least an interval scale. This is accomplished by expressing each item p-value in terms of within-group deviations of a normal curve (see Guilford, 1954, pp. 418-419).

The distance of an item point to the line can be treated as a measure of the degree of item bias. One can determine which items are "greatly deviating" from the line by incorporating outlier or residual analysis. One method is to place confidence limits on the line by using a multiple of the standard error of estimation. An alternate approach, adopted by Strassberg-Rosenberg and Donlon (1975) and Hicks, et al., (1976) involves computing the standard deviation of the residuals and classifying as biased those items deviating by greater than 1.5 standard deviation units. Rudner (1977c) has employed a fixed item-regression line distance of .75 z-score units.

An example of the approach is shown in Figure 1. The transformed p-values have a correlation of approximately .90, making the plot relatively long and

flat. The solid line represents the main axis and the dotted lines represent linear confidence limits. The item represented in the upper left, outside the confidence interval, would be considered biased.

Chi-Square

This approach to biased item analysis determines whether examinees of the same ability level have the same probability of a correct response regardless of cultural affilation. This is accomplished by dividing the tryout samples into groups based on their observed score and comparing the proportions of students within each level responding correctly with a chi-square test for independent observations (Scheuneman, 1975, 1976; Green and Draper, 1972). An item is considered unbiased if, for all individuals in the same total score interval, the proportion of correct response is the same for both groups under consideration. A modified chi-square test determines the probability that an item is unbiased by this definition.

Scheuneman (1976), in applying the approach to several sets of data, advocates using four or five total score levels based on the score distribution of the smaller sample (Green and Draper had used within-group quintiles).

Item Characteristic Curve Theory

Latent trait or item characteristic curve (icc) theory relates the probability of a correct item response to a function of an examinee's underlying ability level (θ_i) and characteristic(s) of the item. While the various models (Lord, 1952; Rasch, 1960; Birnbaum, 1968; Urry, 1970) differ in terms of the number of item parameters considered; they all describe the item parameter(s) independently of the examined sample. Full development of these and other mental measurement models can be found in Hambleton and Cook (1977).

This modern measurement theory has been used to identify biased items (Green and Draper, 1972; Pine, 1976; Lord, 1977; Rudner, 1977a). In an early study, Green and Draper (1972) had used observed total scores as estimates of examinees' abilities, θ_i's, and the proportions of examinees responding correctly at each total score level as estimates of $P(u_g=1|\theta_i)$. Their procedure called for plotting estimates icc's for each item separately for each culture group and comparing the plots.

By this and other latent trait theory approaches, an item is unbiased if examinees of the same ability level, but of different cultural affiliations, have equal probabilities of responding correctly. That is, an item is unbiased if the estimated icc's obtained from the various culture groups are identical. As an example of a biased item, consider the two hypothetical curves shown in Figure 2. These curves are based on responses by two different culture groups to the same item. Total observed scores are used as estimates at θ_i and proportions of examinees responding correctly are used as estimates of $P(u_g=1|\theta_i)$. The curves are not identical, since the location parameters for the two curves are not equal. Such an item can be considered biased in that often examinees of the same ability level, e.g. X_j = 58%, but from different culture groups, do not have similar proportions of correct responses. While this approach is appealing, total observed scores are directly incorporated and quantification of the degree of item bias is difficult (an eyeballing procedure is used to identify a "very biased item").

Rather than using total observed scores as estimates of θ_i and proportions as estimates for $P(u_g= 1|\theta_i)$, more accurate values can be obtained using one of the recent methods of parameterization (Urry, 1975; Wingersky and Lord, 1973).

During parameterization, the metric used for the θ scale is defined by the ability variance in the examined sample. In order to compare parameters obtained from two different examinee groups, the obtained values must be equated. Lord and Novick (1968, Chapter 16.11) have shown that this can be accomplished by computing the regressions of the parameter values based on one group of examinees on the parameter values based on the other group of examinees.

Rudner (1977a) has refined the procedure used by Green and Draper to identify biased items by incorporating equated icc parameter values. The area between pairs of equated icc's is used to indicate the relative amount of aberrance for each item and eyeballing of the equated icc's is employed to provide additional information as to the nature of the aberrance.

Factor Score

In factor analysis, underlying factors (i.e., dimensions or traits) are hypothesized and the correlations of each variable with the hypothesized factors are computed. In an achievement test, each item is treated as a variable. Such an analysis could be conducted twice using examinees from two different cultural backgrounds. Ideally, the two separate groups of examinees would yield similar sets of item-trait correlations (factor loadings). Different sets of factor loadings would indicate that the two groups are not responding to the items in the same manner. Such a test would be considered biased in that it appears to measure a different trait across groups. The items exhibiting the most bias would then be those with the largest differences in factor loading.

Merz (1973, 1976a) has suggested an approach which incorporates factor scores and analysis of variance. In this approach, the item responses for the groups are combined, factor analyzed, and factor scores for each examinee on each factor computed. These factor scores are then subjected to an analysis of variance, with group membership being the independent variable.

Where significant mean differences are found in factor scores, the factor is classified as biased. Biased items are defined as those with high factor loadings on a biased factor.

METHOD

Item Sample

The 1973 Stanford Achievement Test, Form A, Primary 2 Battery, Reading Comprehension Subtest (SAT), -- which, item for item is equivalent to the Stanford Achievement Test - Hearing Impaired Version, Level 2, Reading Comprehension Subtest — formed the item pool for use in this study.

The SAT consists of 16 paragraphs with a total of 48 four-choice items. According to the test publishers, the Psychological Corporation, reading vocabulary is geared to the primary grade levels and emphasis is placed on comprehending disconnected discourse. It was anticipated that the SAT would contain several items biased in favor of one of the incorporated culture group samples.

Examinee Samples

Item responses made by large samples from two diverse culture groups were used in the study. The first culture group was composed of 2,637 students in programs for the hearing impaired across the United States. The scores on the SAT for this group were approximately normally distributed with a mean of 21.6 and a standard deviation of 7.42. This culture group was divided into two subgroups by randomly assigning the examinees to one of two independent groups with significantly different ($p<.01$) mean total scores. Both subgroups were approximately normally distributed. The first subgroup contained 1,079 examinees with a mean of 23.7 and standard deviation of 7.43. The second subgroup contained 1,030 examinees with a mean of 20.9 and a standard deviation of 6.97. Since the examinees were from the same culture group, the expected degree of aberrance for each item was zero. That is,

the approaches were expected to be insensitive to the differential performance of the examinee groups and consistently identify item aberrance as minimal.

The second culture group, representative of the population for which the SAT was designed, was composed of 1,607 examinees from a large west coast public school system. This scores on the SAT for this hearing group were bimodally distributed with modes at 15 and 44, and mean of 28.9 and 12.44.

One major difference between these two culture groups is their exposure to, and their ability to use, the English language (see Stoke, 1976 for an excellent discussion on the social and cultural characteristics of the hearing impaired). Thus, aside from cultural differences, the two groups of examinees greatly differed in their mean level of ability as measured by total score on the SAT.

Procedures

The degree of bias for each item within the SAT was identified by applying a select approach within the transformed item difficulties, icc theory, factor score and chi-square categories to item responses made by (1) the two diverse culture group samples, and (2) two equal culture group samples.

Each item bias detection approach was applied to item responses made by these culture group pairs in the following manner:

transformed item difficulties -- Two sets of item p-values were computed for each culture group pair and transformed to within group normal deviates. From the bivariate scatterplot of the sets of transformed p-values, the absolute values of the magnitudes of the item residuals, i.e. the item-45 degree line distances, were computed. This residual magnitude served to indicate the relative amounts of item bias.

icc theory -- Two sets of item icc parameters as defined by Birnbaum's three parameter logistic model were estimated for each of the SAT items by

separately applying the Urry (1975) iterative minimium chi-square procedure to the item responses of each of the two culture groups. The parameter value estimates were then equated by computing the between group linear regressions for the difficulty and discrimination parameters. The areas between estimated equated icc's, as approximated by:

$$\phi_g = \sum_{-5.000}^{5.000} [| P(u_g=1|\theta_i) - P'(u_g=1|\theta_i) |] \Delta\theta_i$$

where $P(u_g=1|\theta_i)$ and $P'(u_g=1|\theta_i)$ define the estimated equated icc's

and $\Delta\theta_i = .005$

served to indicate the extent of item aberrance.

factor score — The item responses on the SAT made by the two culture groups within each pair were combined and inter-item product-moment correlations computed. The resultant matrix was then reduced using principal component factor analysis with an eigenvalue criterion of 1.0. The factor matrix was rotated orthogonally (varimax) to simple structure and factor scores for each examinee on each factor computed. Separate t-tests were computed using each set of factor scores as dependent variables and group membership as the independent variable. Factors for which there were significant ($p<.001$) differences between mean culture group factor scores were classified as biased. The magnitude of the factor loading (λ_{gj}) on such factors served as indicators of the magnitude of item bias. ϕ_g was then defined as the maximum item factor loading on factors classified as biased. That is,

$$\phi_g = \max [\lambda_{gj}] \qquad j = 1, 2, 3 \ldots \text{number of biased factors}$$

chi-square — Each item was tested individually for bias using a modified

chi-square technique with $i = 2$ culture groups and $j = 5$ total score intervals. By this approach, the expected values for each cell (E_{ij}) were obtained by multiplying (1) the proportion of all examinees with total scores within interval j responding correctly to the item by (2) the number of examinees within the cell. That is,

$$E_{ij} = \frac{O_{.j}}{N_{.j}} (N_{ij}) \qquad i = 1, 2 \qquad j = 1, 2, 3, 4, 5$$

where $O_{.j}$ is the number of examinees in total score interval j responding correctly

$N_{.j}$ is the total number of examinees in interval j

N_{ij} is the total number of examinees in Group i and score score interval j.

As with a conventional chi-square, observed cell values were simply the number of examinees within the cell responding correctly to the item. For each item, the magnitude of aberrance was indicated (1) by the value of the resultant X^2 and (2) by one minus the probability associated with the X^2.

Statistical Analysis

Statistical and graphic analysis were conducted to obtain a global perspective of the similarities and differences among the methodologies. The following analyses were employed:

1. The relative amount of similarity between pairs of approaches was determined by respective Pearson Product-Moment correlations.

2. The identified degrees of bias were compared, item by item, by examining graphs in which items are represented on the abscissa and degree of item bias on the ordinate.

Results

Diverse Culture Group Comparison

The indices of aberrance for each approach to biased item identification

for the diverse culture group comparison are given in Table 1. In the ICC approach, two items, 21 and 44, could not be parameterized because of near zero item-test correlations, and hence could not be evaluated. Seven factors with eigenvalues exceeding unity were extracted by the principal components analysis and rotated orthogonally. Significant differences ($p < .001$) between the mean factor score for the two culture groups were found for six factors. Table 1 shows the maximum factor loading for each item on one of these six factors. The values for the Transformed Item Difficulties ranged from .04 to 1.25.

Because of the dissimilar total score distributions, a problem was encountered in applying the chi-square approach. Initially, five observed score intervals were defined for each item according to the number of examinees in the hearing sample that responded correctly to the item. This resulted in highly disproportionate numbers of hearing impaired examinees in each interval. Also, defining intervals based on the item response distributions of the hearing impaired examinees resulted in highly disproportionate numbers of hearing examinees in each interval. A compromise was achieved by averaging the proportions of examinees responding correctly to the item of each observed score levels across groups, and using four intervals instead of five.

In addition to using the X^2 value to indicate the relative amount of aberrance, one minus the probability associated with the chi-square was used. Both indices are included in Table 1. The use of the probability value as an index identified 56 percent of the items in the SAT as substantially aberrant at $(1-p) > (1-.001)$.

The correlations between the indices of aberrance for each method in the diverse culture group comparisons are given in Table 2. The chi-square - ICC (.67) and the chi-square - transformed item difficulties (.59) correlations were significant at $p<.01$. All correlations involving the chi-square and transformed item difficulties approaches were significant indicating some degree of similarity between each of these approaches and the other models. The factor score and chi-square (1-p) approaches showed the lowest degree of similarity with the other approaches. The average correlation of each of these with the other approaches was .29 and .25, respectively; while the average correlation with other approaches for the chi-square (X^2), transformed item difficulties, and ICC approaches were .48, .37, and .36, respectively.

Equal-culture Group Comparison

The indices of aberrance for the item responses in the equal-culture group comparisons for each approach are given in Table 3. The transformed item difficulties correlated highly ($r = .98$) and all the perpendicular item main axis line distances were minimal. The maximum distance was .28. No items would appear to be identified as biased by this approach.

In the icc approach, again items 21 and 44 did not fit the model and could not be evaluated. Items 28 and 39 showed the most aberrance with values of .51 and .74, respectively. Both of these items showed less aberrance in the diverse culture group comparisons indicating possible misclassification by this approach.

Fourteen factors with eigenvalues exceeding unity were extracted by the principal components analysis and rotated orthogonally. Significant differences ($p<.001$) between the mean factor scores for the two equal-culture groups were found for three factors. The maximum factor loading for items on these three factors ranged between .06 and .72. This range is about the same as the range noted in the diverse culture group comparisons.

Using the chi-square approach, five total score intervals were defined based on the average proportions of examinees responding correctly. The chi-square values obtained were considerably smaller than the values obtained in the diverse culture group comparisons, and no items would have been classified as aberrant at the .05 level.

Figure 3 gives a plot of the aberrance indices for each item for each approach in the diverse culture group comparison and the equal-culture group comparison. It is apparent from Figure 3 that for each approach the variance of aberrance in the equal-culture group comparison is less than the diverse culture group comparison. In the euqal-culture group comparisons, both the factor score approach and the chi-square (1-p) approach appear to have an undesirable amount of variation.

DISCUSSION

The diverse culture group comparison illustrated the approaches as they might be applied in actual test development. Large numbers of examinees from two different populations responded to a pool of items purported to measure the same ability - reading comprehension. Each approach identified a degree of item aberrance for each item. The results show that there was some agreement in terms of the identified degrees of aberrance between (1) the transformed

item difficulties and chi-square (magnitude) approaches and (2) the icc theory and chi-square (magnitude) approaches, although the agreement was not overwhelming ($r = .59$ and $r = .67$, respectively). One minus the probabilities associated with the X^2's and the factor score approach showed little agreement with any of the other methodologies.

Whether the identified degrees of aberrance are in agreement has little direct meaning in test development. A more pertinent question is: Do the approaches lead to the same decisions with regard to which items to classify as "very biased"? If the answer were in the affirmative, the most appealing approach would be the simplest one. Table 4 illustrates which items would be classified as "very biased" by the icc theory, transformed item difficulties and chi-square (magnitude) approaches under the following decision rules:

(a) icc theory - area $\geq .50$

(b) transformed item difficulties - distance $\geq .60$

(c) chi-square (magnitude) - $X^2 \geq 65.0$

These decision rules were determined by identifying, from Figure 3 cut-points which appear to define outliers. Since the variances of the identified degrees of aberrance for the factor score and chi-square (probabilistic) approaches were small, any reasonable cut-point would have resulted in large numbers of items being classified as "very biased" thus these approaches are not included in the table.

From Table 4, it is apparent that the approaches, under these decision rules, would have commonly identified items 16, 17, and 22 as "very biased." Two approaches would have identified items 4, 15, 18, 26, 27, 30

and 45 as being biased. Items 8, 23, 24, 25, 29, 44 and 47, however, were identified by only one approach. More conservative or more liberal decision rules would still have resulted in different sets of items being identified.

Since there is some disagreement among the approaches, the results of the equal-culture group comparison warrant closer examination. The two groups of examinees in this comparison were from the same well-defined population; namely, students with a hearing loss sufficient enough to warrant a special educational program. As such, item bias between these two groups is by definition minimal, and the expected amounts of aberrance identified for each item by each approach is assumed to be zero.

Of the approaches, only the transformed item difficulties approach fully met this criterion. The identified degrees of aberrance from this approach were small, and by any reasonable decision rule, no items would have been classified as biased. Thus, the model behaved as expected. The identified degrees of item aberrance as indicated by the icc theory approach were also minimal. However, two items could not be evaluated and two items would have been identified as having fair amounts of aberrance under a liberal decision rule.

The icc theory approach unexpectedly identified items 28 and 39 as containing fair amounts of bias. A closer examination of these items reveals that their latent trait item difficulty parameters were extreme for the second group of examinees, namely 2.77 and 3.91 respectively. This can be loosely interpreted as meaning that, ignoring guessing, an examinee's ability must be 2.77 (3.91) standard deviations above the group mean ability to have a better than average chance of responding correctly. Since relatively few examinees were of this ability level, parameterization became tenuous and the slight aberrance in these items is probably due to abnormally high parameterization

error. Thus, this approach is liable to yield spurious results when item difficulty is extremely high or low. It should be noted that the number of items in the SAT is really insufficient for a proper evaluation of the icc approach. From a Monte Carlo investigation of the Urry parameterization procedure, Schmidt and Gugel (1975) have recommended that a minimum of 60 items and 1,000 subjects be used to obtain accurate parameter estimates. Since the SAT contains only 45 items, the parameter value estimates may have contained more than the usual amounts of error.

Items 21 and 44 had extremely low item-test point biserial correlations, which implied that ability was poorly related to the probability of a correct response. Such items cannot fit the Birnbaum model and hence cannot be evaluated for bias with the icc theory approach. Although such items are usually the first to be eliminated in test development, the fact that these items cannot be evaluated illustrates a weakness in the approach.

The chi-square approach in the equal-culture group comparison produced wide fluctuations in the probabilities associated with the X^2's used to test the null hypothesis of no bias. However at $p<.05$, $[(1-p)>.95]$, no items were suspected as being biased. Thus, although 56 percent of the items were identified as biased in the diverse-culture group comparison, in terms of the equal-culture group comparison, the chi-square approach appeared to be sufficient when either probabilities or magnitudes were employed.

The factor score approach identifies aberrant items as those having a major loading on a factor which yields unequal mean factor scores. In the equal-culture group comparison, three sets of mean factor scores were identified as unequal at conservative values ($p<.001$). The maximum loadings of many items on these factors were high, several being higher than the maximum loading in the diverse culture group comparison. The approach, as applied to the data

in this study, produced unsatisfactory results in the equal-culture group comparison.

The above discussion has pointed out that there were differences between the approaches in the identified degrees of aberrance in both the diverse-culture group and equal-culture group comparisons. Of the methodologies, the transformed item difficulties and icc theory approaches appear most attractive. In the diverse-culture group comparison several items were identified as biased, and in the equal-culture group comparison, the identified degrees of aberrance were minimal. The factor score approach did not identify much variance in item bias in the diverse-culture group comparison and yielded major loadings in the equal-culture group comparison. Using a conservative probability level ($p<.001$) the chi-square approach identified 56 percent of the items as biased in the diverse culture group comparison and yielded wide fluctuations in the amount of aberrance in the equal-culture group comparisons.

These later two approaches - the chi-square approach and the factor square approach - both incorporate significance testing of large amounts of data. The chi-square approach examines the hypothesis that the proportions of examinees responding correctly are identical across individuals in the same observed score interval and of different cultural classifications. The factor score approach incorporates the hypothesis that the group mean factor scores are identical across the defined culture groups on each factor. With samples as large as that used in this study, hypothesis testing may not be appropriate. The sample values are such that they can be considered population values and small differences are statistically significant.

In the diverse-culture group comparison, the X^2 values correlated with the distances of the transformed item difficulties approach and the areas of the icc theory approach. However, their magnitudes were extreme. It should

be noted that in the diverse culture group comparison, the total score distributions of the examinee samples were quite divergent. In the equal-culture group comparison, the distributions were not as different and the X^2 values were substantially less.

The chi-squrre approach analyzes the item response data in terms of observed score intervals. The observed value for an interval and culture group is simply the number of examinees in the interval and culture group responding correctly to the item. The expected value for a culture group and interval is the product of proportion of all examinees in the interval responding correctly to the item and the number of examinees in the culture group and in the interval. Thus, the expected value will be influenced by the culture group with the greater number of examinees in the interval when the observed score distributions are different. Since the item interval definitions are often similar, this will result in a near systematic inflation of the X^2 values.

An example of how total score distributions affect the expected interval values (and consequently the X^2 values) is illustrated by the hypothetical item response data shown in Table 5. Here, the total observed score distributions are quite different. Group 1 has more than five times as many examinees in the interval as does Group 2. Further, the total number of examinees at each total score level within the interval decreases as total score increases for Group 1 and increases for Group 2. However, the proportions of examinees responding correctly to the item at each total score level are identical across groups. That is, the two groups perform identically within the interval and their total score distributions are dissimilar. If the approach were not

sensitive to total score distributions, the observed and expected values for each group would be identical. However, the observed and expected values are:

$$\text{for group 1,} \quad O_1 = 136 \text{ and } E_1 = \frac{136 + 31}{480 + 90} \cdot 480 = 140.6, \text{ and}$$

$$\text{for group 2,} \quad O_2 = 31 \text{ and } E_2 = \frac{136 + 31}{480 + 90} \cdot 90 = 26.4$$

Thus, even though the two groups performed identically at each total score level, the observed and expected values are unequal and would have inflated the X^2 value. Had different distributions been employed, different expected values and a different X^2 would have been defined.

The factor score approach entails many decision points which will affect the results. In this study, phi-correlations of the combined data, principal component analysis, eigenvalues greater than 1.0, varimax rotation, and probabilities less than .001 were used, and the results appeared to be unsatisfactory. In the diverse culture group comparison 26 out of 48 items had a maximum factor loading of .55 ± .10 on a factor yielding significantly different mean factor scores, and the identified degrees of aberrance in the equal-culture group comparison fluctuated widely with several items being identified as being more aberrant than the most aberrant item in the diverse-culture group comparison.

The factor score approach attempts to identify items which most strongly measure traits in which the groups differ significantly. In large scale investigations, groups are likely to differ on any measured trait including the ones intended by the test publisher and those unintentionally built into the test. Thus, a significant difference in the mean factor scores on the main test factor may be of little interest. Differences on other factors,

however, would indicate the presence of items which inappropriately influence group mean scores. In order to identify these items, the underlying factors of the test must be well-defined and the major factor clearly identified. Principal component analysis using eigenvalues greater than one and varimax rotation does not appear to allow for this. Principal component analysis yields factors which are defined by the data (as opposed to inferred), a unity eigenvalue criteria does not guarantee that the correct number of factors will be extracted and varimax rotation can obfuscate the major factor. A different set of factor analytic procedures might have yielded more equitable results.

It should be noted that the factor score approach incorporates a definition of item bias which is substantially different than the other approaches. The approach seeks to identify items which measure a trait other than that measured by the remaining items of the test (by factor analyzing the combined data) and heavily contribute to differential performance (by contributing to differential mean factor scores). Generically, the other approaches are concerned with which items measure different traits across groups and operationally with which items behave differently across groups. This distinction is not as subtle as it may appear. The other approaches are incapable of identifying items which measure a trait other than that gauged by the other items when the groups perform equitably.

The two more attractive approaches, the transformed item difficulties and the icc theory approaches, also incorporate different operational definitions of bias. The transformed item difficulties approach identifies items which, relative to the other items in the test, are more difficult for members of one group than they are for members of another group of examinees. The icc theory approach identifies items for which examinees of the same true ability and from different population groups have unequal probabilities of a

correct response. Thus, the transformed item difficulties approach addresses aggregate group performance as indicated by item p-values and the icc theory approach addresses the range of item performance along the ability continuum as indicated by item characteristic curves.

The difference between these two approaches is illustrated by items 25 and 17 (in Figure 4). In the diverse culture group comparison, item 25 was identified as biased by the icc theory approach and not by the transformed item difficulties approach. The overall difficulty of the item for the two diverse-culture groups was about equal. Consequently, the item was not identified by the transformed item difficulties approach. However, low ability hearing impaired examinees and high ability hearing examinees are favored. That is, when considered across ability levels the item behaved differently between groups. Item 17, which was identified by both approaches, does not show this type of inverted differential performance. Across the ability continuum, hearing examinees are favored.

When comparing the transformed item difficulties and icc theory approaches in terms of different decision rules, five items were commonly identified by both approaches. All five of these items were of this latter type - noninverted differential performance across the ability continuum. This further illustrates that the transformed item difficulties approach is sensitive to differences in mean item difficulty while the icc theory approach appears to be sensitive to both mean item difficulty and to group performance along the continuum. However, it should be noted that different definitions of item difficulty, and hence mean group performance, are employed. The transformed

item difficulties approach directly defines item difficulty from the aggregate data. The icc theory approach infers item difficulty from performance on the item alone. Since these different definitions are employed, different items were identified as being biased against a group as a whole.

Conclusions

Based on the two applications, the factor score and chi-square approaches appeared to be inadequate for identifying biased items. The X^2 values in the chi-square approach were shown to become inflated as total observed score distributions differ, thus leading to erroneous classifications of bias. The factor score approach, which incorporates a somewhat different definition of bias, identified large degrees of aberrance in the equal-culture group comparison. It was felt that the decisions used in factor analyzing the data led to the unsatisfactory results. It was further noted that both of these approaches employed inference testing which may not be appropriate with the large sample sizes used in this study.

The transformed item difficulties and the icc theory approaches appeared to be most promising. The identified degrees of aberrance in the equal-culture group was consistently low for both approaches, although a liberal decision rule would have led to the false identification of one or two items by the icc theory approach. The two approaches identified several items in common in the diverse culture group comparison. The major difference between these two methodologies is that the icc theory approach appears to be sensitive to bias against both individuals and groups of examinees and the transformed item difficulties approach appears to be sensitive to bias only against groups.

Recommendations

The investigation utilized a single set of diverse culture group data

for which the item parameters were unknown a priori. While there was substantial reason to suspect the presence of some biased items, the true number of biased items, their amounts of aberrance and their item numbers were unknown. A similar study using simulated data with known parameters may prove revealing. Such a study could also investigate the behavior of the approaches under different numbers of biased items.

One of the more promising and interesting approaches to the detection of biased items, the distractor response analysis (Veale and Foreman, 1975, 1976; Maw, 1977), was not evaluated in this study - due to the lack of the appropriate item response data. Rather than analyzing the numbers of examinees responding correctly, this approach identifies differences in distractor response patterns. Although the approach incorporates inference testing, it may prove beneficial to the field and should be considered in future investigations of item bias detection methodologies.

REFERENCES

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. New York: College Entrance Examination Board, 1973.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968, Chapts. 17-20.

Cardall, C. & Coffman, W. R. A method for comparing performance of different groups on the items in a test. (RM 64-61) Princeton: Educational Testing Service, 1964.

Cleary, T. A., & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Darlington, R. B. Another look at "cultural fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. Intelligence and Cultural Differences. Chicago: University of Chicago Press, 1951.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75(4), 261-269.

Green, D. R., & Draper, J. F. Exploratory studies of bias in achievement tests. Monterey: CTB/McGraw-Hill, 1972.

Gross, A. L., & Su, W. Defining a fair of unbiased selection model: a question of utilities. Journal of Applied Psychology, 1975, 60, 345-351.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Hambleton, R. K. & Cook, L. L. Latent trait models and their use in the analysis of educational data. Journal of Educational Measurement, 1977, 14(2), 75-96.

Hicks, M. M., Donlon, T. F. & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished Doctoral Dissertation, University of Washington, 1972.

Jensen, A. P An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: Eric Clearinghouse, 1973, (ERIC Document Reproduction Service ED 086 726).

Lord, F. M. A theory of test scores. Psychometric Monograph Number 7. Princeton: Educational Testing Service, 1952.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores (2nd Ed.). Reading, MA: Addison-Wesley, 1968.

Lord, F. M. A study of item bias using item characteristic curve theory. Proceedings of the Third Congress of Cross-Cultural Psychology, Tilburg, Holland, 1977.

Maw, C. E. Item response patterns and group differences: an aplication of the log-linear model. Unpublished doctoral dissertation, University of Chicago, 1977.

Merz, W. R. Factor analysis as a technique in analyzing test bias. Paper presented at the annual meeting of the California Educational Research Association, Los Angeles, 1973.

Merz, W. R. Estimating bias in test items utilizing principle component analysis and the general linear solution. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Merz, W. R. Test fairness and test bias: a review of procedures. In M. Wargo and D. R. Green Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. New York: McGraw-Hill, 1977, in press.

Pine, S. M. "Application of Item Characteristic Curve Theory to the Problem of Test Bias," in Weiss, D. J. (Ed.). Applications of Computerized Adaptive Testing, Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October, 1976.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Denmarks Paedoggiogishe Institute, 1960.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977a.

Rudner, L. M. Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, June 1977b.

Rudner, L. M. Item Bias with Deaf and Hearing Examinees. Volta Review, 1977c, in press.

Schmidt, F. L., & Gugel, J. F. The Urry Item Parameter Estimation Technique: How Effective? Paper presented at the American Psychological Association Convention, Chicago, August 1975.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975.

Scheuneman, J. A procedure for evaluating item bias in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Stokoe, W. C. The study and use of sign language, Sign Language Studies, 1976, 10, 1-36.

Strassberg-Rosenberg, B., & Donlon, T. F. Context influences on sex differences in performance and aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Thorndike, R. L. Concepts of cultural fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. Unpublished Doctoral Dissertation, Purdue University, 1970.

Urry, V. W. Ancillary estimators for the parameters of mental text models. Paper presented at the American Psychological Association Convention, Chicago, August 1975.

Veale, J. R., & Foreman, D. I. Cultural validity of items and tests: A new approach. Score Technical Report, Iowa City, Iowa: Westinghouse Learning Corporation/Measurement Research Center, 1975.

Veale, J. R., & Foreman, D. I. Cultural variation in criterion-referenced tests: a "gobal" item analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Williams, R. L., Mosby, D., & Hinson, V. Critical issues in achievement testing of children from diverse ethnic backgrounds. In M. Wargo and D. R. Green, Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. New York: McGraw-Hill, 1977, in press.

Wingersky, M. S., & Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses. (RM 73-2.) Princeton: Educational Testing Service, 1973.

TABLE 1

Degrees of Aberrance Identified by the Approaches in the Diverse-Culture Group Comparison

Item #	ICC Area	Transformed Item Difficulties	Chi Square (1-p)	Chi Square (X^2)	Factor Score
1	.40	.24	.98	5.9	.35
2	.07	.31	.999	33.1	.53
3	.29	.13	.87	8.5	.55
4	.75	.79	.999	54.2	.45
5	.25	.21	.99	11.1	.61
6	.17	.18	.89	6.2	.40
7	.15	.43	.99	11.9	.45
8	.50	.54	.999	27.9	.46
9	.27	.14	.99	12.6	.35
10	.24	.46	.99	11.1	.42
11	.34	.54	.999	42.8	.62
12	.37	.52	.999	43.6	.60
13	.11	.52	.999	55.1	.52
14	.16	.05	.60	3.0	.28
15	.25	.68	.999	105.4	.42
16	.57	1.11	.999	107.7	.61
17	.76	1.25	.999	159.0	.65
18	.83	.85	.999	27.7	.26
19	.37	.23	.999	30.7	.30
20	.16	.18	.99	14.8	.36
21	-	.44	.99	14.4	.56
22	2.30	.67	.999	240.9	.52
23	.38	.67	.999	31.8	.23
24	.61	.51	.98	10.2	.53
25	1.01	.08	.999	49.5	.57
26	.38	.67	.999	94.8	.60
27	.04	.76	.999	65.2	.48
28	.32	.18	.96	8.2	.55
29	.29	.44	.999	65.4	.34
30	.23	1.05	.999	122.3	.52
31	.13	.07	.999	26.0	.36
32	.19	.01	.65	4.2	.27
33	.14	.15	.99	13.7	.44
34	.15	.05	.96	8.2	.17
35	.14	.66	.999	17.7	.33
36	.09	.17	.999	33.6	.22
37	.07	.32	.18	.9	.26
38	.14	.43	.999	34.7	.20
39	.23	.14	.99	15.1	.36
40	.08	.37	.999	23.4	.44
41	.27	.16	.60	2.8	.51
42	.27	.33	.60	2.9	.46
43	.07	.16	.999	22.8	.46
44	-	.26	.999	133.2	.48
45	.55	.04	.999	85.1	.49
46	.25	.26	.99	13.4	.51
47	.60	.21	.88	6.1	.57
48	.34	.24	.999	33.1	.44

Table 2

Correlations of the Degrees of Aberrance Identified by the Approaches in the Diverse Culture Group Comparison

	Transformed item difficulties	Chi-Square (X^2)	Chi-Square (1-p)	Factor score
Icc theory	.31*	.67**	.17	.28
Transformed item difficulties		.59**	.29*	.30*
Chi-square (X^2)			.31*	.34*
Chi-square (1-p)				.23

* p .05
** p .01

TABLE 3

Degrees of Aberrance Identified by the Approaches in the Equal-Culture Group Comparison

Item #	ICC Area	Transformed Item Difficulties	Chi-Square (1-p)	Chi-Square (X^2)	Factor Score
1	.12	.02	.32	2.4	.19
2	.15	.02	.22	1.7	.07
3	.10	.16	.05	.5	.16
4	.06	.06	.32	2.4	.36
5	.08	.18	.01	.1	.07
6	.28	.14	.48	3.3	.06
7	.24	.09	.08	.9	.26
8	.19	.03	.01	.2	.02
9	.19	.08	.52	3.4	.32
10	.08	.02	.28	2.1	.09
11	.18	.00	.03	.5	.19
12	.17	.11	.28	2.1	.14
13	.04	.13	.01	.2	.19
14	.21	.12	.12	1.2	.20
15	.04	.07	.18	1.6	.26
16	.22	.03	.40	2.6	.13
17	.31	.15	.48	3.3	.20
18	.26	.07	.08	.9	.57
19	.32	.03	.68	4.8	.20
20	.24	.04	.15	1.4	.46
21	-	.28	.68	4.7	.11
22	.17	.05	.40	2.6	.06
23	.34	.14	.06	.7	.15
24	.19	.21	.09	1.0	.20
25	.36	.09	.68	4.8	.08
26	.21	.01	.03	.6	.17
27	.11	.02	.07	.8	.40
28	.51	.16	.59	3.8	.14
29	.11	.09	.26	2.0	.40
30	.14	.14	.53	3.7	.19
31	.09	.10	.12	1.1	.14
32	.07	.03	.31	2.3	.24
33	.34	.12	.78	5.6	.25
34	.14	.13	.20	1.7	.72
35	.12	.21	.73	5.3	.70
36	.22	.18	.07	.8	.72
37	.06	.15	.26	2.1	.63
38	.23	.09	.48	3.3	.34
39	.74	.16	.88	7.6	.10
40	.38	.06	.47	3.2	.20
41	.35	.14	.81	6.5	.08
42	.37	.05	.52	3.5	.11
43	.29	.08	.12	1.2	.11
44	-	.12	.31	2.3	.08
45	.34	.06	.48	3.4	.48
46	.14	.10	.07	.8	.51
47	.32	.07	.08	.9	.68
48	.26	.16	.68	4.8	.43

TABLE 4

Items classified as biased (***) by three approaches under select decison rules in the diverse-culture group comparison

ITEM #	ICC THEORY	TRANSFORMED ITEM DIFFICULTIES	CHI-SQUARE (X^2)
1	-	-	-
2	-	-	-
3	-	-	-
4	***	***	-
5	-	-	-
6	-	-	-
7	-	-	-
8	***	-	-
9	-	-	-
10	-	-	-
11	-	-	-
12	-	-	-
13	-	-	-
14	-	-	-
15	-	***	***
16	***	***	***
17	***	***	***
18	***	***	-
19	-	-	-
20	-	-	-
21	-	-	-
22	***	***	***
23	-	***	-
24	***	-	-
25	***	-	-
26	-	***	***
27	-	***	***
28	-	-	-
29	-	-	***
30	-	***	***
31	-	-	-
32	-	-	-
33	-	-	-
34	-	-	-
35	-	***	-
36	-	-	-
37	-	-	-
38	-	-	-
39	-	-	-
40	-	-	-
41	-	-	-
42	-	-	-
43	-	-	-
44	-	-	***
45	***	-	***
46	-	-	-
47	***	-	-
48	-	-	-

Table 5

Hypothetical Item Response Distributions by Total Score Levels Within a Single Interval

		N in each total score level		N responding correctly	
		Group 1	Group 2	Group 1	Group 2
total score level	10	200	10	40 (20%)	2 (20%)
	11	160	30	48 (30%)	9 (30%)
	12	120	50	48 (40%)	20 (40%)
		480	90	0_1 = 136	0_2 = 31

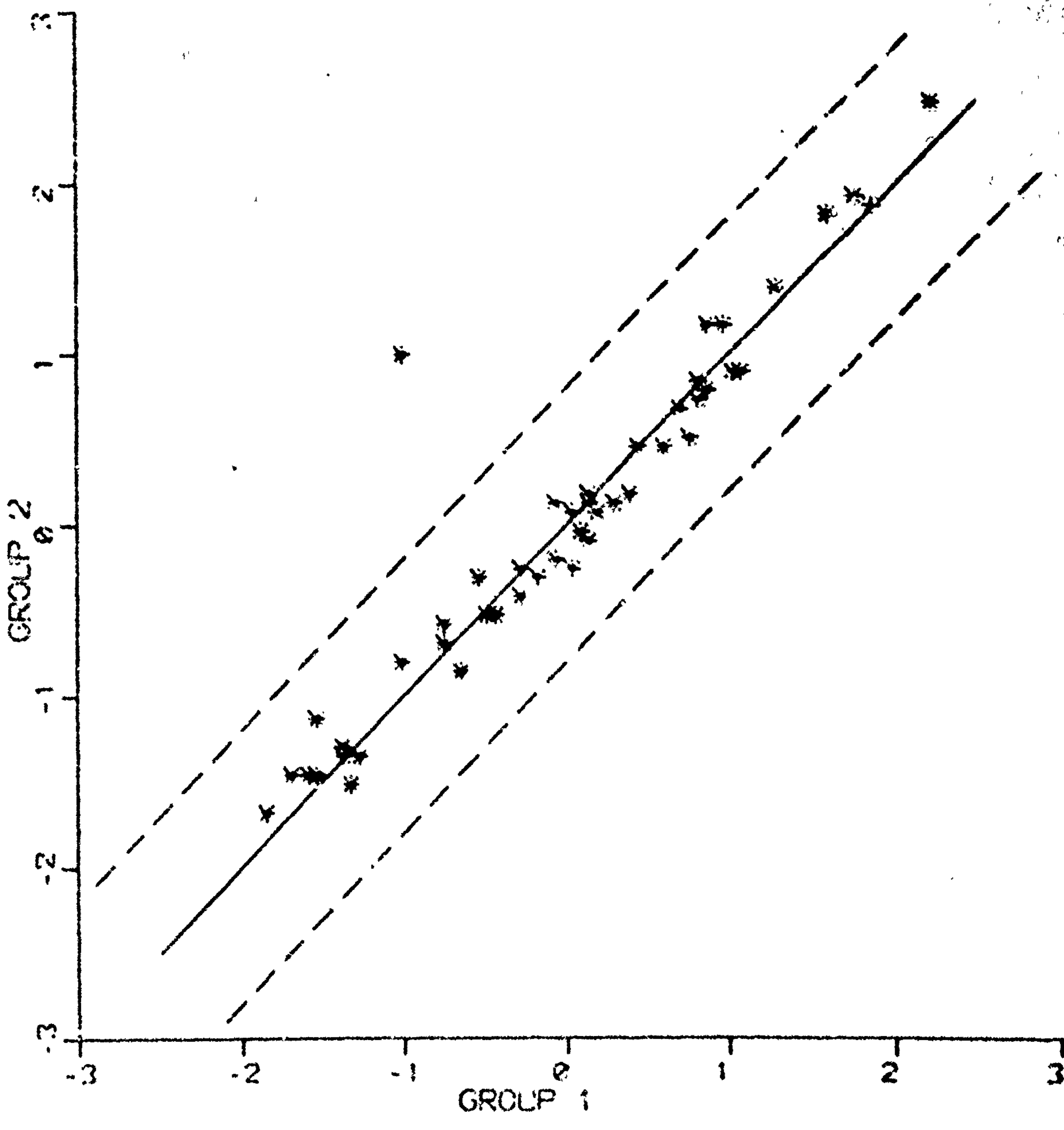


Figure 1: A Hypothetical transformed item difficulties Scatterplot

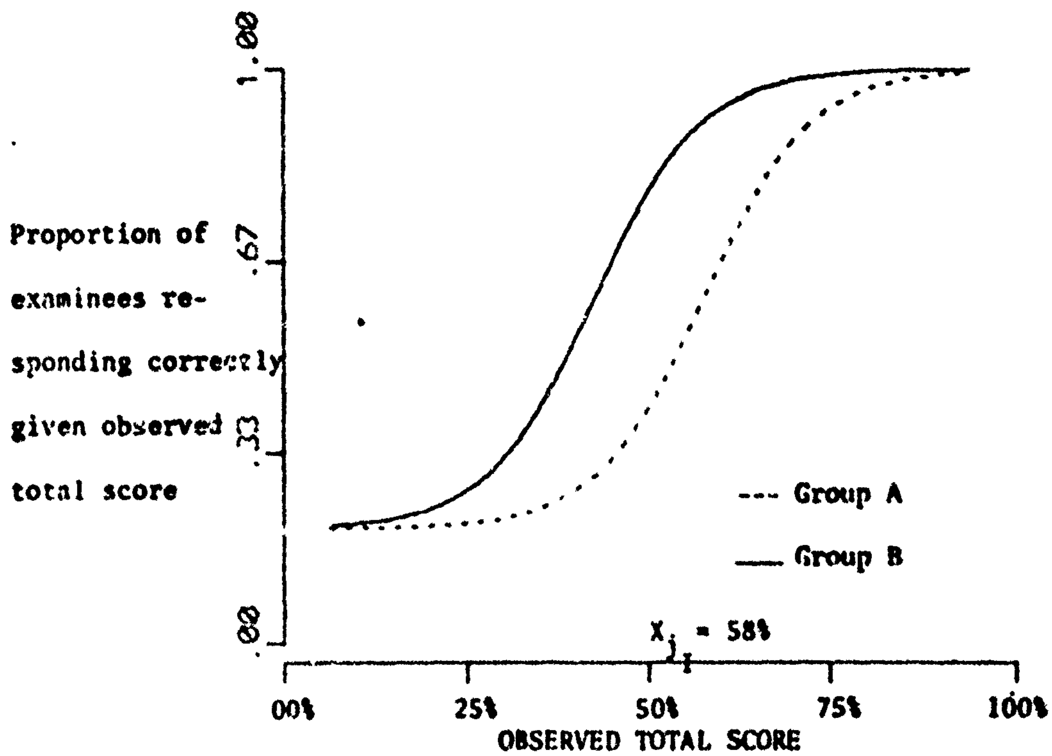


Figure 2: Two hypothetical response distributions

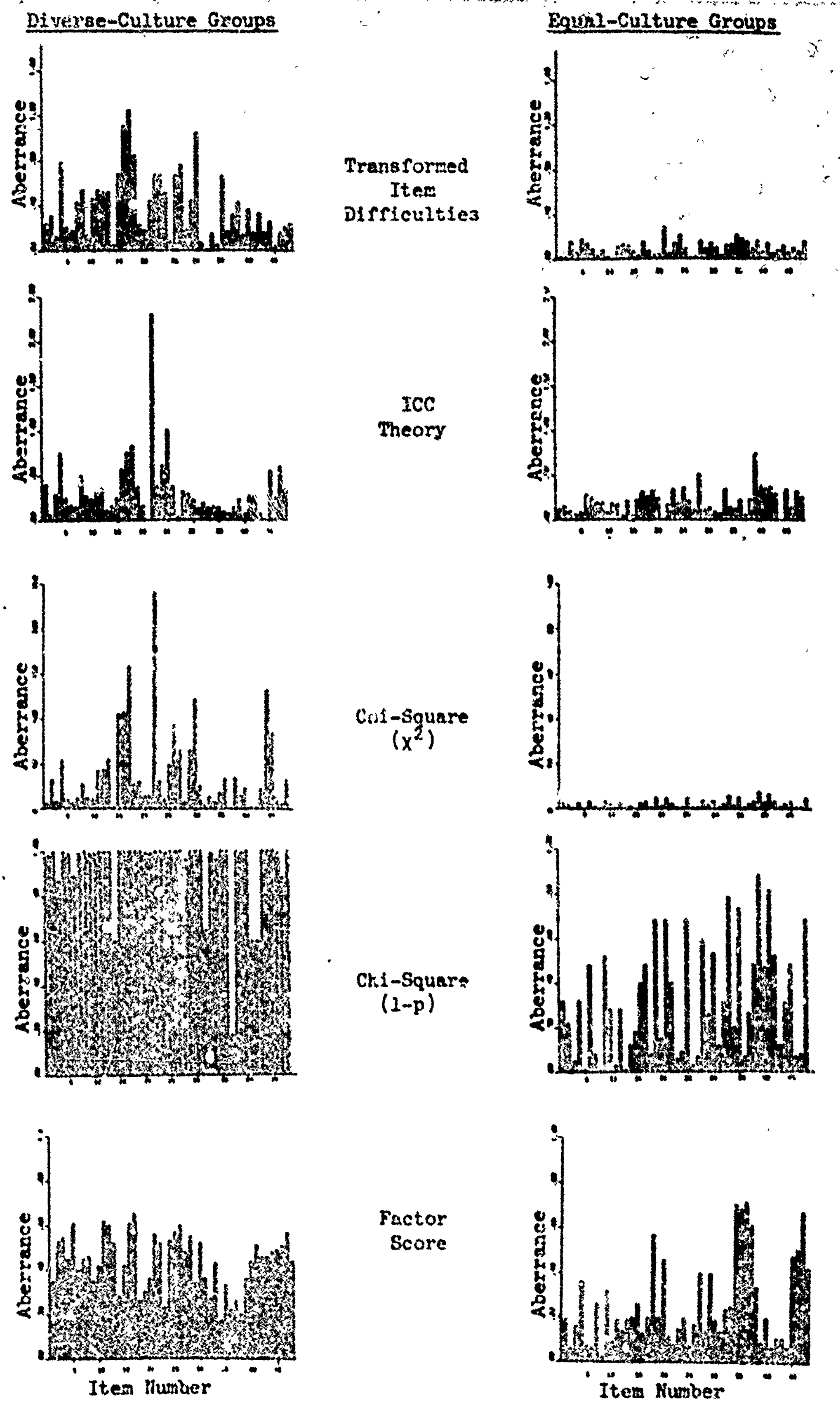
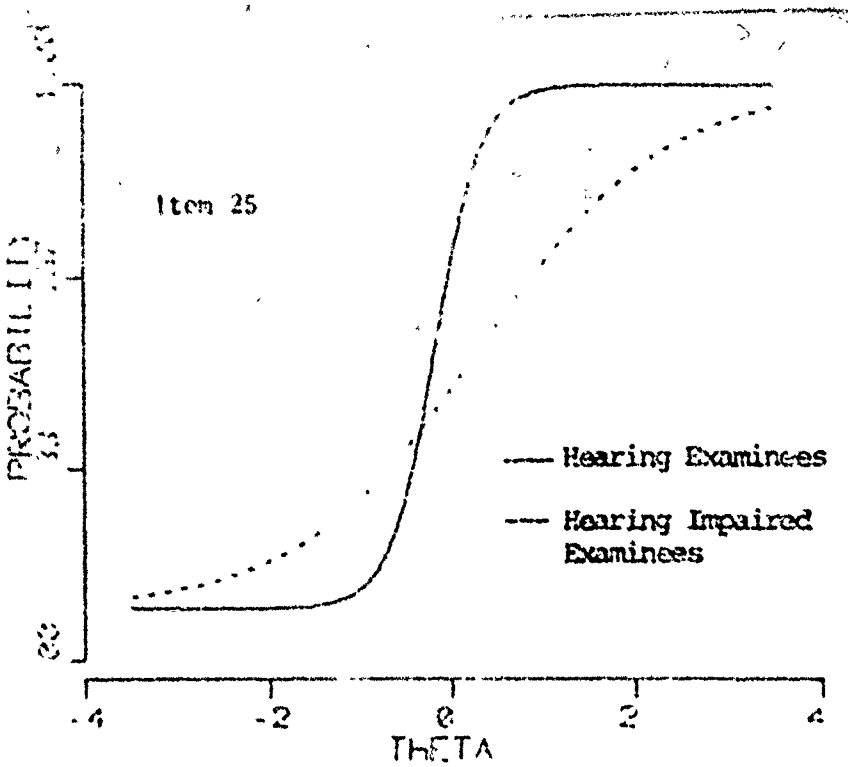


Figure 3: Plots of the degrees of aberrance identified by each approach for each group comparison.



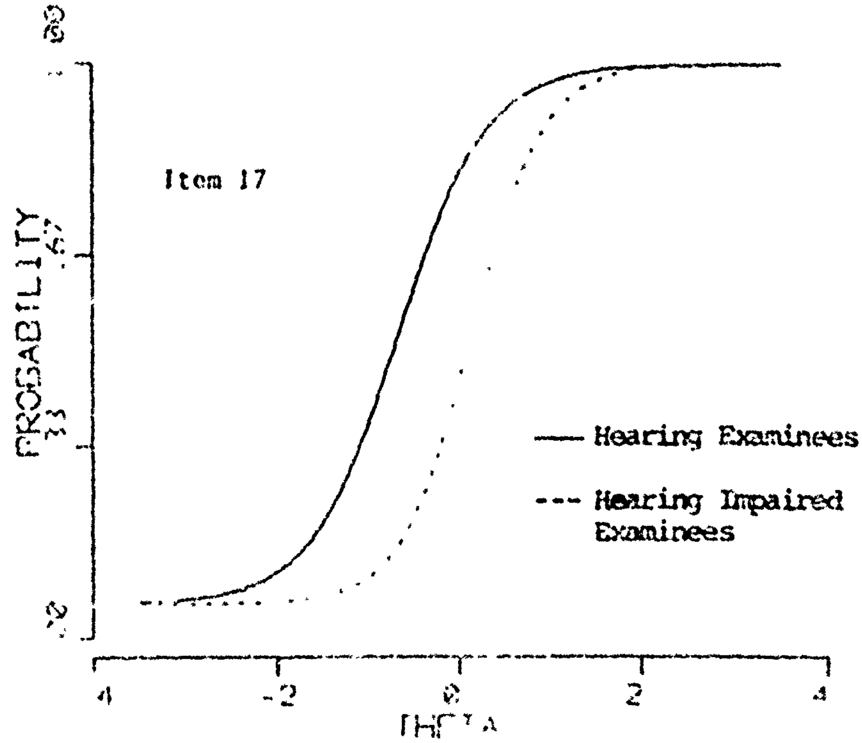


Figure 4: Estimated equated icc's for items 17 and 25 in the diverse-culture group comparison.

A STUDY OF THE UTILITY OF LATENT TRAIT SCORING FOR MEASURING THE CULTURE-FAIRNESS OF TESTS

By
Charles H. Cory
Navy Personnel Research and Development Center

Personnel with submarginal literacy have been a continuing problem to the Navy and over the years have been processed by means of a variety of special programs designed to identify and either upgrade or eliminate them. At the present time the Navy is administering a reading comprehension test early on in Boot Camp to about 30 percent of incoming recruits in order to identify the small percentage of submarginal readers in the enlisted input. On the basis of the test results, recruits who are found to have a reading grade level (RGL) below 3.0 are given administrative discharges; those having RGLs between 3.0 and 5.4, inclusive, are assigned to Academic Remedial Training; and those with RGLs of 5.5 or better are assigned directly to Recruit Training.

Reading Grade Level is a good predictor of recruit attrition versus non-attrition for personnel who are reading in the submarginal range. Studies at the Navy Personnel Research and Development Center (NPRDC) have found that 64 percent of recruits with RGLs less than 4.0 attrite prior to completion of Recruit Training. This compares with 20 percent Boot Camp attrition for personnel reading in the 4.0-5.9 RGL range and 10 percent attrition for personnel reading in the 6.0-7.9 RGL range. Recently an effort was undertaken to determine whether the relationship of reading grade level to attrition from Recruit Training was substantially the same or whether it was substantially different for Blacks and Whites.

To examine this question, records for all Whites and all Blacks were extracted from a large, full-range sample of incoming recruits who had been given the Gates-MacGinitie Test of Reading Comprehension, Survey D, grades 4-6. Two-by-two tables, formed separately for the White and Black subsamples were designed to show the disposition of personnel with RGLs less than 4.0 and greater than or equal to 4.0 relative to attrition/nonattrition in Recruit Training.

Table 1 shows these data for Whites. You can see that reading score has a moderately high correlation with non-attrition in Recruit Training. As would be expected, the great majority of the group reading above the fourth grade level did not attrite in Recruit Training and the phi coefficient of .29 indicates that a moderately high positive relationship existed between reading score and non-attrition in Recruit Training. From the left hand column of Table 1 it is apparent that more Whites with RGLs below 4.0 attrited in Recruit Training than did not attrite in Recruit Training.

As is shown in Table 2, for Blacks the relationship between reading score and attrition in Recruit Training was somewhat less than for Whites, (phi coefficient of .16 compared with .29) but it was still significantly different from zero at $p < .001$. Correlations of these magnitudes in the two racial groups appear to meet the requirements for unbiased tests which have been specified in the proposed Uniform Guidelines on Employee Selection Procedures of the Equal Employment Opportunity Coordinating Council.

However, from the left hand column of Table 2, it can be seen that about one Black in four with less than a fourth grade reading level actually attrited from Recruit Training. In fact, only 26 percent of Blacks compared with 57 percent of Whites with RGLs below 4.0 attrited in Recruit Training. These differences are disturbing, and they indicate that subtle biases, particularly at the item level, may be present in the test. Accordingly, it was decided to investigate if a more

sophisticated scoring technique using a latent trait model could detect bias in the test. The study which was undertaken in this connection had the objectives of (1) detecting item bias in the test if it existed and measuring its magnitude and (2) evaluating the utility of a latent trait model for reducing the differences in the screen rates of Blacks and whites.

Latent trait or item characteristic curve(ICC) models, relate the probability of success on an item to a function of the trait being measured and to the characteristics of the item. A frequently used model for ICCs and one which was used for the present study is a 3-parameter logistic model. This model assumes, that (1) an examinee's probability of responding correctly on an item is unaffected by his response on any other item (i.e. the items have local independence) and (2) that all items in the test measure a single trait.

Figure 1, an example of an ICC for a multiple choice item with five alternatives, relates the probability of answering the item correctly (on the ordinate) to the ability of the examinee, shown on the abscissa. Ability is normally denominated by the Greek letter, θ.

You can see that the probability of a correct response is a monotonic function of θ. At sufficiently low levels of ability, represented by 1 1/2 standard deviations below the mean, the probability of answering this item correctly is essentially at a chance level. In contrast, for very high levels of ability, the probability of getting the item correct approaches unity.

A formula for the 3-parameter logistic model, which would be used to determine the probability of a correct response for any given θ to item i, is shown at the bottom of Figure 1. The term $P_i\theta$ is the probability that an individual with ability θ will correctly answer item i; D is a constant scaling factor, and a_i, b_i, and c_i are parameters of i which together describe the relationship between ability level and probability of a correct response.

b_i, the item difficulty parameter, indicates the level of θ which, excluding the effects of guessing, has exactly a 50 percent probability of answering i correctly. For the curve in Figure 1, b_i would be about 0. a_i is proportional to the slope of the item characteristic curve measured at the point on the ICC which is directly above b_i. It indicates the rapidity at which the probability of a correct response changes with changes in ability level. Since it measures the ability of the item to discriminate among levels, it is called the discrimination parameter. c_i is the lower asymptote of the ICC for item i and represents the probability that a person having no knowledge will guess the correct answer to i. Notice that c_i, the empirical probability of guessing correctly, may or may not be the same as the theoretical probability of guessing the correct answer. For a 5-choice item shown, the theoretical probability of guessing correctly is the reciprocal of the number of choices for the item, or .20. The item parameters a_i, b_i, and c_i and the ability parameter, θ, are measured on the same scale and are typically expressed as standard scores.

Extensive research has been conducted with latent trait models for scoring tests. Tests created using these models have been found to produce substantial savings in administration time and/or to be associated with greater accuracy of measurement than similar tests which are administered using traditional formats and methods of scaling.

The ability to develop tests which are free of item bias has also been claimed as an advantage of latent trait scoring methods. In fact, Petersen (1977) and Lord (1977) among others, have pointed out that latent trait techniques represent the only adequate means for evaluating bias at the item level. In the latent trait model an item would be considered unbiased if the probability of getting it correct is the same for all examinees of a given ability, regardless of group membership.

Despite the theoretical attractiveness of the latent trait appreach, very few studies of item bias have made use of this technique. Furthermore, the major studies

using latent trait scaling to evaluate item bias have either employed simulated data to test the characteristics of the model (Pine & Weiss, 1976) , or have been focussed on identifying items whose ICCs for Blacks and Whites were statistically identical, except for sampling variation (Lord, 1977). There appear to be no studies in the literature which consider the differences in ICCs of racial groups in relation to scores on external criteria.

The Navy Personnel Research and Development Center has a computer program developed by Dr. Urry of the U. S. Civil Service Commission which uses a 3-parameter logistic model to scale test items and to compute theta values. For the present research it was decided to apply this program to samples of Blacks and Whites extracted from a full range sample of more than 30,000 incoming recruits who had been administered the Gates-MacGinitie Reading Test (GM), Survey D, grades 4-6 in connection with research conducted by Dr. Duffy of NPRDC. Records of personnel in this data base also contain a code for attrited/non attrited in Recruit Training, the AT code.

Complete samples of Blacks and Whites who took form 2 of the GM were extracted from the data base and were used to create two separate Black-White data sets. The Truncated (TR) data set consisted of personnel who had scores of 50 or lower on the General Classification Test, a test of verbal ability. In other words, the TR file was composed of personnel having verbal abilities in a range that, in the Navy, is routinely screened for reading ability. All blacks and every seventh White having GCT scores in this range were included in tho TR data set--a total of 680 Blacks and 733 Whites.

The Full Range (FR) data set consisted of personnel having GCT scores across the entire range. Thus it included the total Black sample and every twelfth case in the total White sample, a total of 951 Blacks and 993 Whites. For the TR and FR data sets statistics on the Gates MacGinitie tests are shown in the next table.

It can be seen from the table that the mean reading scores for Blacks were about one half and one fourth standard deviations below those for the Whites in the FR and TR samples respectively and the standard errors of the test were somewhat larger for Blacks than for Whites. In contrast, the standard deviations and KR 20 coefficients of Blacks and Whites were substantially similar. These findings are not very surprising; in fact, they are consistent with the usual findings for scores of Blacks and Whites on paper-and-pencil tests.

At the present time, comparison of potential racial bias in item characteristic curves has a kind of stopgap or seat-of-the-pants quality because the distribution characteristics of the item parameters are not known. Consequently, there is no specific test which is generally accepted as appropriate to evaluate the statistical significance of Black-White differences in ICCs--as, for example the Gulliksen-Wilks test is appropriate for evaluating the statistical significance of differences in linear regression statistics of Blacks and Whites.

In the Urry program a 2-step process is used to develop item parameters. The first estimates of the parameters are based on standardized total test scores and the second are computed using Bayesian modal estimates of ability. The latter values are basically estimates obtained by applying the first estimated parameters, $\hat{a}$, $\hat{b}$, and $\hat{c}$, to the scored items for each examinee to compute an estimate of individual ability, $\hat{\theta}$. In turn, the probability of getting the item correct is plotted against these $\hat{\theta}$s and revised $\hat{\hat{a}}$, $\hat{\hat{b}}$, and $\hat{\hat{c}}$ parameters are computed. During the process of computing double-hat parameters, cases for which stable estimates of $\hat{\hat{\theta}}$ cannot be computed are dropped. Also, during this step, items with $\hat{\hat{a}}$, $\hat{\hat{b}}$, and $\hat{\hat{c}}$ values which have poor characteristics are excluded from the calculation of $\hat{\hat{\theta}}$.

Thus the $\hat{\hat{\theta}}$ ability estimates have been purified of internal inconsistencies in two respects: (1) they are not based on scores from items with ICCs which have

poor characteristics and (2) they do not contain unstable $\hat{\theta}$ values. For these reasons $\hat{\theta}$ might be expected to be more accurate than RGL as an estimate of ability. In addition, the $\hat{\theta}$ values in the present study were based on ICCs which were specific to the individual racial group. If test bias were a cause of the differential in the Black-White screen rates of RGLs, comparisons based on $\hat{\theta}$ values computed in this fashion would be expected to decrease the differences. Based on these considerations, the following sets of comparisons were carried out for both the TR and the FR data sets.

1. Counts were made of the items, categorized as shown in the next slide.
2. For the items in categories 1 and 2 on the slide, the a, b, and c parameters of Blacks and Whites were formed into bivariate distributions and inspected for similarities and differences.
3. Two-by-two tables relating scores on $\hat{\theta}$ and AT and RGL and AT were formed and inspected.
4. AT was regressed on $\hat{\theta}$ and on RGL separately for Blacks and Whites.
5. $\hat{\theta}$ values were correlated with RGL and other cognitive measures as well as with an additional two performance criteria which were available on the records.

<u>Results</u>

For both Blacks and Whites the percentage of items for which stable $\hat{a}$, $\hat{b}$, and $\hat{c}$ and $\hat{\hat{a}}$, $\hat{\hat{b}}$, and $\hat{\hat{c}}$ parameters could be computed together with the percentages of individuals in the samples who had stable $\hat{\theta}$ values are shown in the next table.

The following relationships are apparent from the table: 1. Stable ICCs could be computed for more of the items for Blacks than for Whites. 2. A drastic reduction in percentages of items having stable values occurred for the computation of the double-hat parameters versus the percentages for the single-hat parameters. 3. In general the number of items for which stable ICCs could be computed was somewhat greater for the TR than for the FR data set, but the differences were not large. 4. The percentages of the samples for which stable $\hat{\theta}$ ability estimates could be computed was slightly larger for Blacks than for Whites.

Thus the statistics look somewhat better for Blacks than for Whites.

For the items for which a, b, and c parameters were computed for both Blacks and Whites, the values for each of six item statistics for Blacks and Whites were correlated. These coefficients are shown in the next table.

It is apparent that the extent of agreement among Blacks and Whites in terms of the item statistics was greater for the TR than for the FR data set. Also, for both data sets, the extent of agreement among the coefficients was lower for the double-hat parameter items than for the single-hat parameter items. It is interesting that the agreement is particularly high for the two item difficulty statistics, the <u>p</u> value of classical test theory and the b parameter of latent trait theory. The extent of agreement for the slope parameters also tends to favor the statistic of classical test theory, the point biserial coefficient, although there is considerable variation in these relationships across the sets of items and in one set, the coefficient for the a parameter is higher than that for the point biserial.

The agreement was greater for the point biserial coefficient than for the biserial coefficient for the single-hat items and this relationship was reversed for double-hat items. The values for the c parameter had substantial correlations for the single-hat parameter items but these correlations were completely eliminated for the double-hat items.

Scattergrams for these statistics for the TR Data Set are shown in the next four figures. For each scattergram, the regression line of Whites on Blacks is drawn as a solid line and a dashed line indicating perfect correlation is shown.

Figures 2 and 6 illustrate the high degree of agreement between the item difficulties for Blacks and Whites which was characteristic of the statistics for both classical and latent trait theory. As is shown by the intercept of Figure 2, on average, p values were about .11 higher for Whites. In contrast, for the b statistic, the differences between the racial groups were almost nonexistent.

Figures 3 and 5 provide graphic illustration of the greater correspondence between the Black-White distributions for the point biserial coefficient than for the a parameter, as was discussed previously. The location of the intercepts indicate that the slopes of the item-test regressions were generally greater for Whites for both the classical and the latent trait measures.

The next table provides comparisons for Blacks and Whites of the correlations of $\hat{\theta}$ and RGL with three criteria, two classification tests and a biographical variable, years of education. It can be seen that the correlations in general are considerably higher for Whites than for Blacks, a finding which is depressingly consistent with those of most studies. In addition, the hoped for increases in the accuracy of prediction of Recruit Attrition from the use of $\hat{\theta}$ estimates of ability, not only did not materialize, but $\hat{\theta}$ was actually not as good as RGL for predicting most of the variables. As a predictor of attrition of Blacks during Recruit Training, the most important comparison, $\hat{\theta}$ did not even have statistically significant validity coefficients, in contrast to the comparable coefficients for RGL which were significant at the .01 level. The same types of relationships were generally characteristic for the other variables in the table. In general, $\hat{\theta}$ was not as good a predictor as RGL for criteria and did not correlate as high as RGL with other test variables and with years of education.

To provide a comparison with the Pearson product moment correlation coefficients computed for RGL and Recruit Attrition shown in the previous table, phi coefficients were computed for Blacks and Whites grouped into two-by-two tables using the RGL and Recruit Attrition categories that were shown in Tables 1 and 2. The results of these analyses are shown in the next table. For both Blacks and Whites the coefficients are generally about two or three points higher than the Pearson <u>r</u>s shown in the previous table. Thus if it were desired to provide a pre-enlistment screen for low readers, the best cut point would be at an RGL of 4.0.

However, the poor performance of $\hat{\theta}$ as a predictor of Recruit Attrition was disturbing. Therefore an additional set of analyses was performed to throw some light on the reasons for this phenomenon.

As is shown in the next table, the two parts of the GM, Vocabulary and Reading Comprehension, were moderately speeded and not everyone finished all of the items. The speededness was somewhat greater for the Reading Comprehension section than for the Vocabulary section, and for the TR than for the FR data sets. For both the FR and the TR data sets, Blacks were considerably slower than Whites. In general, for any number of items omitted, the percentage of the Black group at that level was from 30 to 200 percent greater than the percentage of the White group at that level.

The next table presents the same number-of-items-omitted comparison; however, the sample for which it was made consists of personnel remaining after computation of the double-hat parameters. You can see that for persons in the refined sample, the number of omitted items has been reduced so that the percentages omitting items at any level is only about half to two thirds of the comparable percentage shown in the previous table. The reduction is particularly drastic for Whites in the FR data set. This indicates that an effect of the refinement process was primarily to eliminate personnel having large numbers of omitted items.

Extrapolations were made from the complete percentage of omissions figures to compute a mean number of items omitted for each of the eight groups. Thus for the FR data set, the mean items omitted by Blacks for the complete and the refined samples were 2.43 and 1.05, respectively. Comparable means for the Whites were 1.05 and .19. For the TR data set the means were 3.27 and 2.21 for Blacks and 2.16 and 1.01 for Whites. Thus, on average Blacks omitted one more item on the test than did Whites and the falloff in mean number of items omitted in the refined sample was about one item for each of the racial groups.

From additional inspections of the data further indications were found that persons eliminated by the computer program during the data refinement process primarily had low reading scores and it was determined that the relationship was consistent across data sets. An indication of the characteristics of the relationship can be seen in the next table, which presents the $\hat{\theta}$ values for personnel in the TR data set classified into personnel having stable $\hat{\theta}$ values and those having unstable $\hat{\theta}$ values. Figures for personnel of the latter type are shown in the M. D. (missing data) column of the table. An indication of the reading ability level of these two different types isshown by the RGL designation in the left hand column of the table. It can be seen that for both Blacks and Whites, in general, personnel having RGLs of 4.0 or lower were dropped from the refined group, personnel reading at fifth and sixth grade levels were sometimes dropped and sometimes had $\hat{\theta}$ values computed, and $\hat{\theta}$ values were computed for all persons reading above the sixth grade level.

Discussion and Conclusions

Whatever the advantages of latent trait scaling may be in terms of economy and accuracy of measurement, in the present study the technique did not serve to provide a better predictor for Blacks and Whites than the commonly used RGL. From the data presented it is clear that several reasons were responsible for the failure. The major one was that, because of characteristics of the program, $\hat{\theta}$ values simply were not available for the personnel having the lowest reading abilities. The restriction of range resulting from this deficiency undoubtedly lowered the Pearson rs computed for $\hat{\theta}$.

However, not all is lost for the latent trait model. For the present research latent trait scoring, even after reducing the number of items between 25 and 50 percent for Blacks and Whites, respectively, provided a measure which substantially reproduced the total score on the complete Gates-MacGinitie test. Even the present set of items, which in terms of their latent trait parameters are considerably below the desirable level, could save a great deal of administration time if they were administered by means of a branching technique which selected each new item to correspond as closely as possible to the estimated ability of the examinee.

A number of comparisons in the study suggest that the item-test relationships were substantially similar for Blacks and Whites. The correlations of $\hat{\theta}$ with RGL were 94 and 90 respectively, for Blacks and Whites. This suggests a high degree of relationship between the $\hat{\theta}$ values and the possibly contaminiated RGL. In addition the item difficulty and slope parameters for the items reviewed indicate a considerable correspondence in these relationships for Blacks and Whites. Although these statistics were undoubtedly for the best items in the test, the high relationship between the RGL and $\hat{\theta}$ suggests that the other items in the test do not substantially modify the correspondences in item statistics of Blacks and Whites.

Indications from inspecting the correlations of the item statistics suggest that greater divergences between Blacks and Whites occur for the latent trait statistics than for the statistics of classical test theory. The determination of which of these sets of statistics most accurately describes relationships in the real world must be left to future research.

The above data suggest that it would be desirable to look in other locations for the explanation of the lower than would be predicted attrition rates for Blacks. Two possible reasons for this phenomenon suggest themselves: (1) Blacks assigned to Academic Remedial Training may simply try harder than Whites assigned to Academic Remedial Training, or (2) because of the current concern with increasing the proportions of Blacks in the Navy and the subsequent reluctance to eliminate Blacks, teaching personnel in ART may either provide more assistance to Blacks than Whites in the training, or may employ more lenient evaluation standards for Blacks than for Whites, or both. Any of these effects would bring about a difference among Blacks and Whites in the strength of the reading level - Recruit Attrition relationship.

REFERENCES

Lord, F. M. A study of item bias, using item characteristic curve theory. Educational Testing Service, Princeton, New Jersey, U.S.A.

Petersen, N. S. Bias in the selection rule -- bias in the test (DRAFT June 1977). Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands.

Pine, S. M., & Weiss, D. J. Effects of item characteristics on test fairness (Research Report 76-5). Department of Psychology, University of Minnesota Minneapolis, MN, December 1976.

TABLE 1

ANALYSIS FOR WHITES AGAINST ATTRITION
IN RECRUIT TRAINING

READING GRADE LEVEL

	<4.0	≥4.0	TOTAL
ATTRITE	307	1427	1734
NON-ATTRITE	233	24,808	25,041
TOTAL	540	26,235	26,775

$r_\phi = \frac{7283565}{24802062} = .29$ $\chi^2 = 26775\ (.29)^2 = 2309.10^{***}$

***p<.001

TABLE 2

ANALYSIS FOR BLACKS AGAINST ATTRITION
IN RECRUIT TRAINING

READING GRADE LEVEL

	<4.0	≥4.0	TOTAL
ATTRITE	10	65	75
NON-ATTRITE	29	1371	1600
TOTAL	39	1636	1675

$r_{\phi} = \frac{13825}{87501.3} = .16$ $\chi^2 = 1675(.16)^2 = 41.81$***

***p<.001

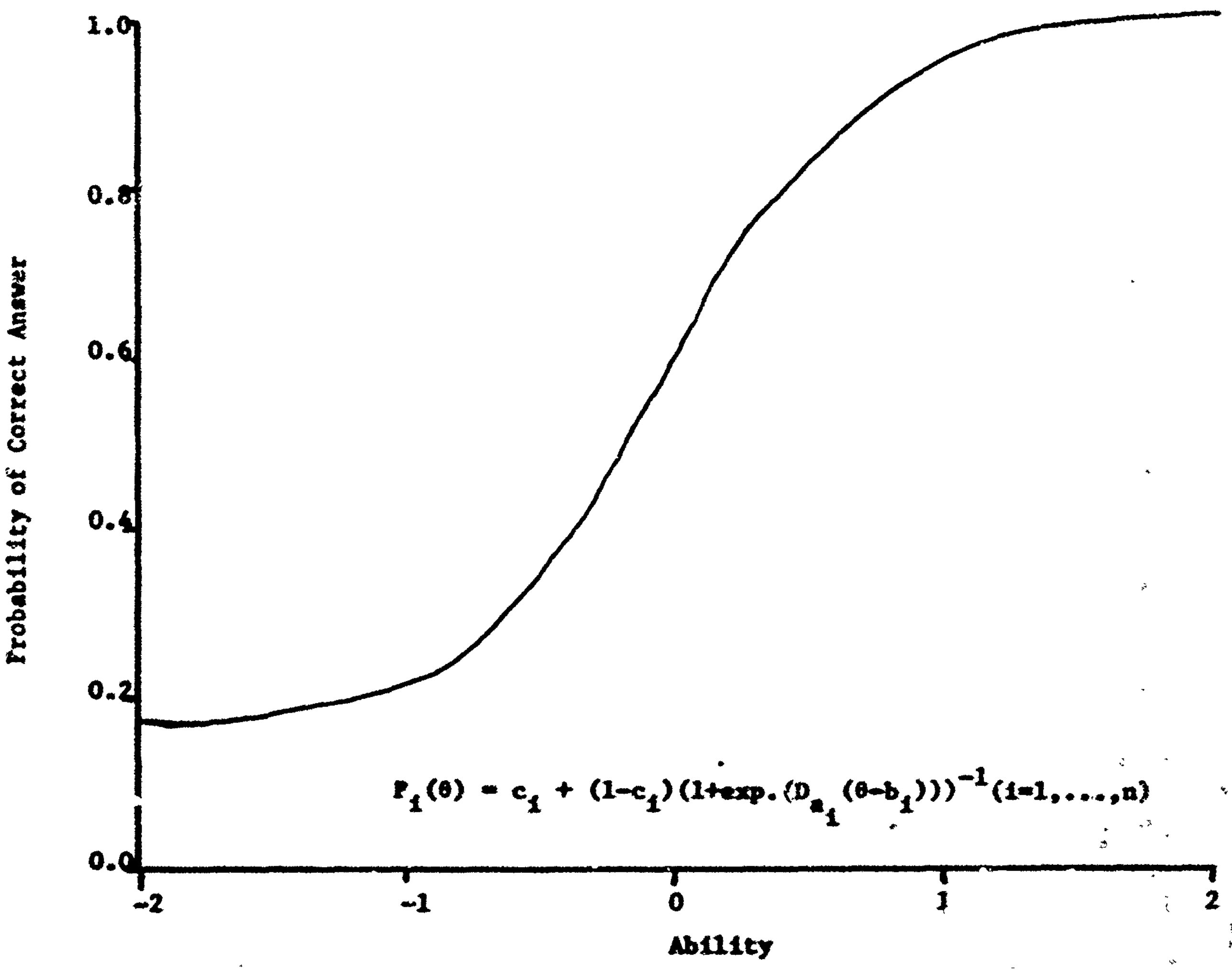


Figure 1: Probability of a correct answer to item i as a function of ability

SAMPLE SIZES AND TEST STATISTICS FOR THE DATA SETS

	N	$\bar{X}$	S.D.	KR 20	S.E.
FR DATA SET					
BLACKS	951	82.9	13.1	.94	3.2
WHITES	993	89.1	12.0	.95	2.7
TR DATA SET					
BLACKS	680	79.0	13.0	.93	3.5
WHITES	733	82.3	13.5	.94	3.2

CATEGORIZATION OF ITEMS ON THE BASIS OF ICC PARAMETERS

1. ITEMS FOR WHICH IN-RANGE $\hat{a}$, $\hat{b}$, AND $\hat{c}$ VALUES WERE COMPUTED FOR BOTH BLACKS AND WHITES
2. ITEMS FOR WHICH IN-RANGE $\hat{\hat{a}}$, $\hat{\hat{b}}$, AND $\hat{\hat{c}}$ VALUES WERE COMPUTED FOR BOTH BLACKS AND WHITES
3. ITEMS FOR WHICH IN-RANGE $\hat{a}$, $\hat{b}$, AND $\hat{c}$ VALUES WERE COMPUTED FOR BLACKS BUT NOT FOR WHITES
4. ITEMS FOR WHICH IN-RANGE $\hat{a}$, $\hat{b}$, AND $\hat{c}$ VALUES WERE COMPUTED FOR WHITES BUT NOT FOR BLACKS
5. ITEMS FOR WHICH IN-RANGE $\hat{\hat{a}}$, $\hat{\hat{b}}$, AND $\hat{\hat{c}}$ VALUES WERE COMPUTED FOR BLACKS BUT NOT FOR WHITES
6. ITEMS FOR WHICH IN-RANGE $\hat{\hat{a}}$, $\hat{\hat{b}}$, AND $\hat{\hat{c}}$ VALUES WERE COMPUTED FOR WHITES BUT NOT FOR BLACKS

COMPARISON OF BLACKS AND WHITES IN TERMS OF ITEMS HAVING ACCEPTABLE ICCs AND CASES WITH STABLE $\hat{\theta}$ VALUES

PERCENTAGES

	TOTAL BLACK	TOTAL WHITE	BOTH BLACK AND WHITE	BLACK NOT WHITE	WHITE NOT BLACK
FR DATA SET					
ITEMS HAVING ACCEPTABLE $\hat{a}$, $\hat{b}$, AND $\hat{c}$ VALUES	76	45	35	41	4
ITEMS HAVING ACCEPTABLE $\hat{\hat{a}}$, $\hat{\hat{b}}$, AND $\hat{\hat{c}}$ VALUES	40	21	19	21	2
CASES WITH STABLE $\hat{\theta}$ VALUES	90	85			
TR DATA SET					
ITEMS HAVING ACCEPTABLE $\hat{a}$, $\hat{b}$, AND $\hat{c}$ VALUES	79	75	66	13	9
ITEMS HAVING ACCEPTABLE $\hat{\hat{a}}$, $\hat{\hat{b}}$, AND $\hat{\hat{c}}$ VALUES	50	30	27	23	3
CASES WITH STABLE $\hat{\theta}$ VALUES	93	90			

BLACK-WHITE CORRELATION COEFFICIENTS FOR STATISTICS FOR ITEMS HAVING URRY PARAMETERS
FOR BOTH BLACKS AND WHITES

DATA SET	$\underline{p}$ value	$\underline{r}$ p. bis.	$\underline{r}$ bis.	a	b	c
<u>FR</u>						
∧ parameter items	.94	.87	.42	.22	.94	.71
⩓ parameter items	.90	.46	.63	.46	.91	-.06
<u>TR</u>						
∧ parameter items	.97	.87	.68	.62	.96	.60
⩓ parameter items	.95	.66	.78	.83	.90	-.02

Product Moment Correlations of $\hat{\theta}$ and RGL for
Written Test, Biographical and Criterion Variables

	FR Data Set				TR Data Set			
	$\hat{\theta}$		RGL		$\hat{\theta}$		RGL	
Variable	Black	White	Black	White	Black	White	Black	White
Successful Completion of Recruit Training	.05	.09**	.09**	.34***	.04	-.06	.10**	.19***
Highest Paygrade Received	.24***	.25***	.26***	.40***	.10**	.01	.16***	.23***
Total Number of Promotions	.00	-.04	.05	.18***	.05	-.05	.08*	.18***
GCT	.72***	.69***	.71***	.68***	.54***	.47***	.59***	.53***
ARI	.51***	.44***	.48***	.47***	.27***	.23***	.29***	.31***
Years of Education	.21***	.30***	.21***	.30***	.05	.09**	.10**	.13***

* $p < .05$

** $p < .01$

*** $p < .001$

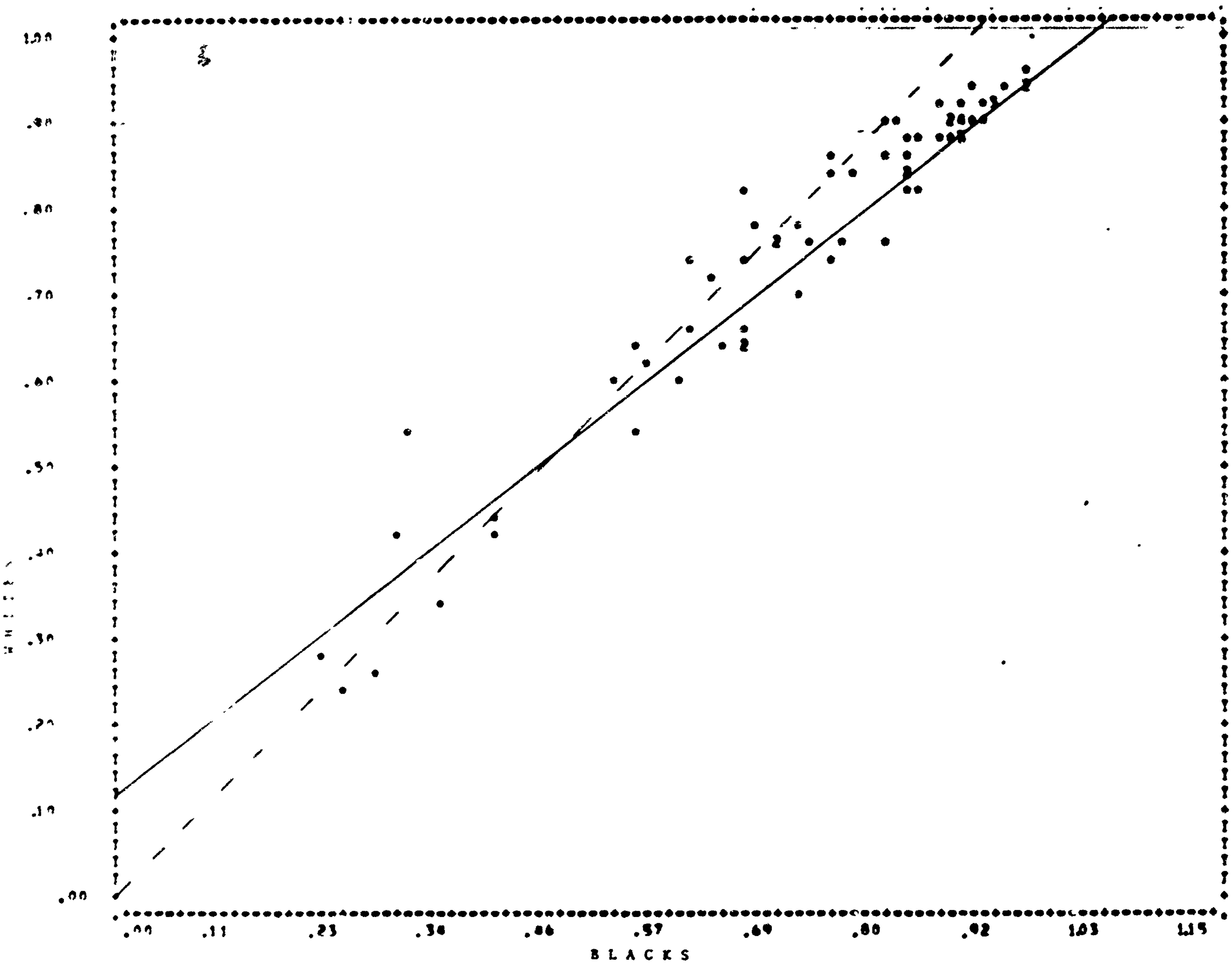
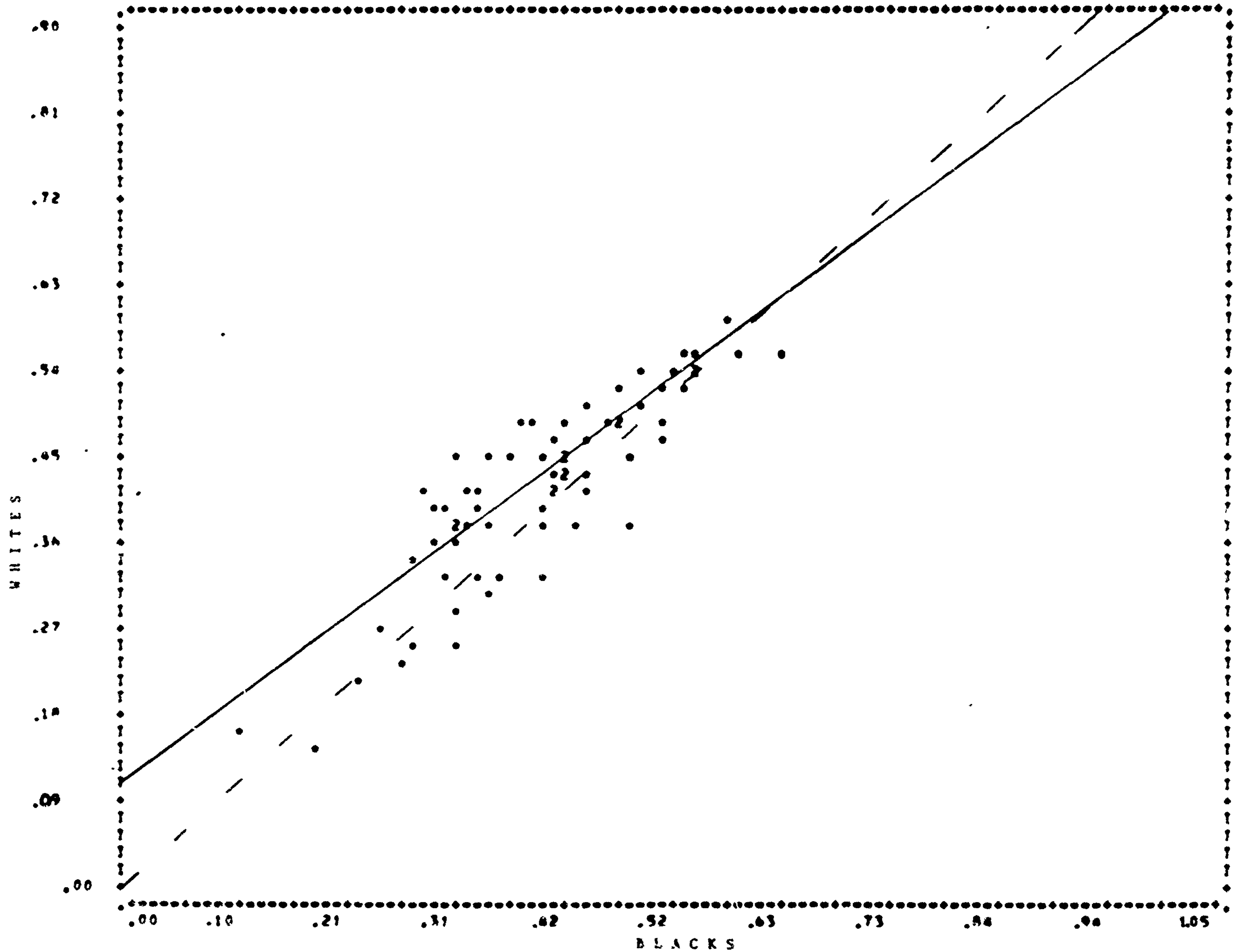


Figure 2. SCATTERGRAM OF THE ITEM P VALUES OF BLACKS AND WHITES.



COEFFICIENTS OF BLACKS AND WHITES

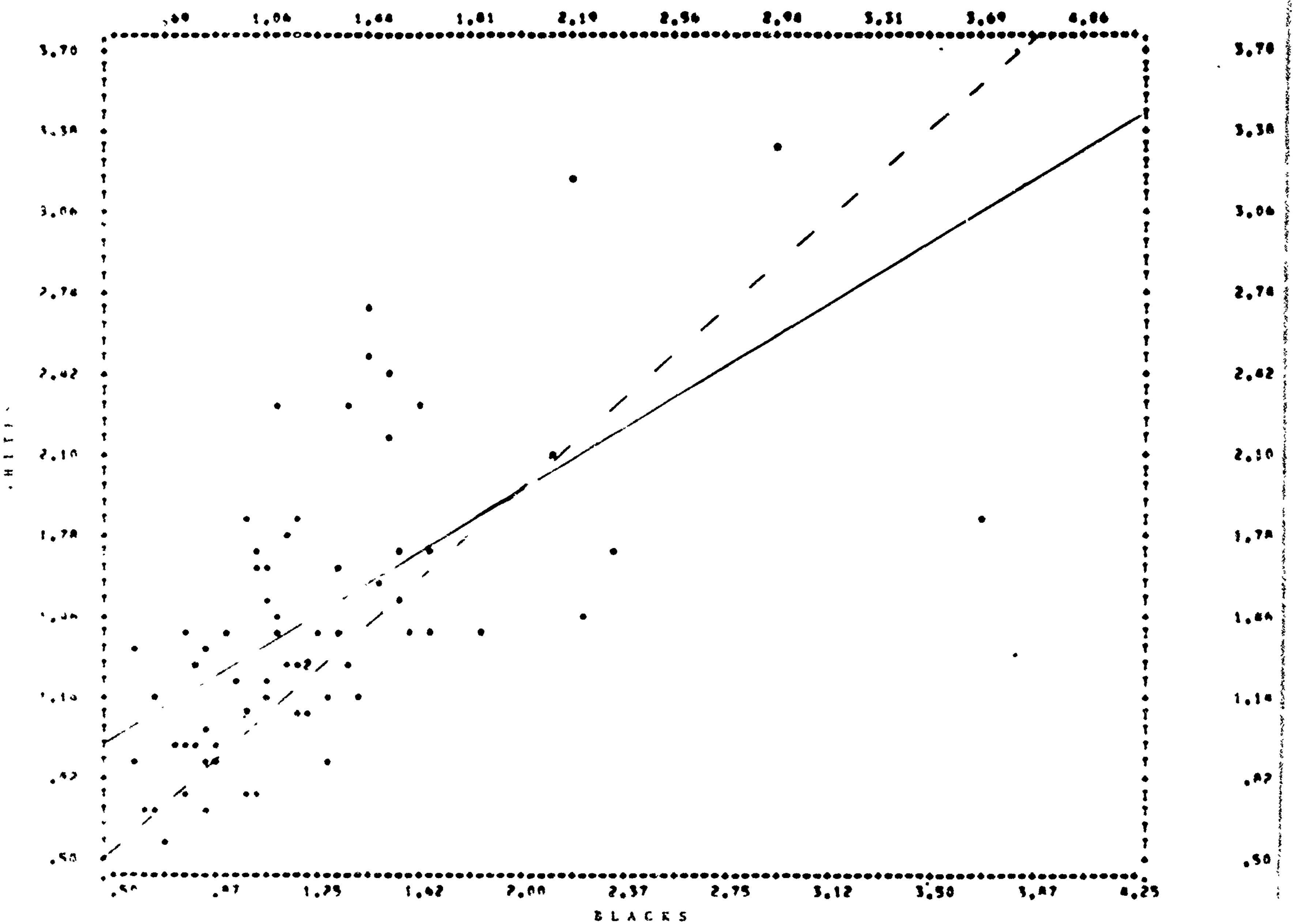


Figure 5. SCATTERGRAM OF THE ITEM [illegible] PARAMETERS OF BLACKS AND WHITES.

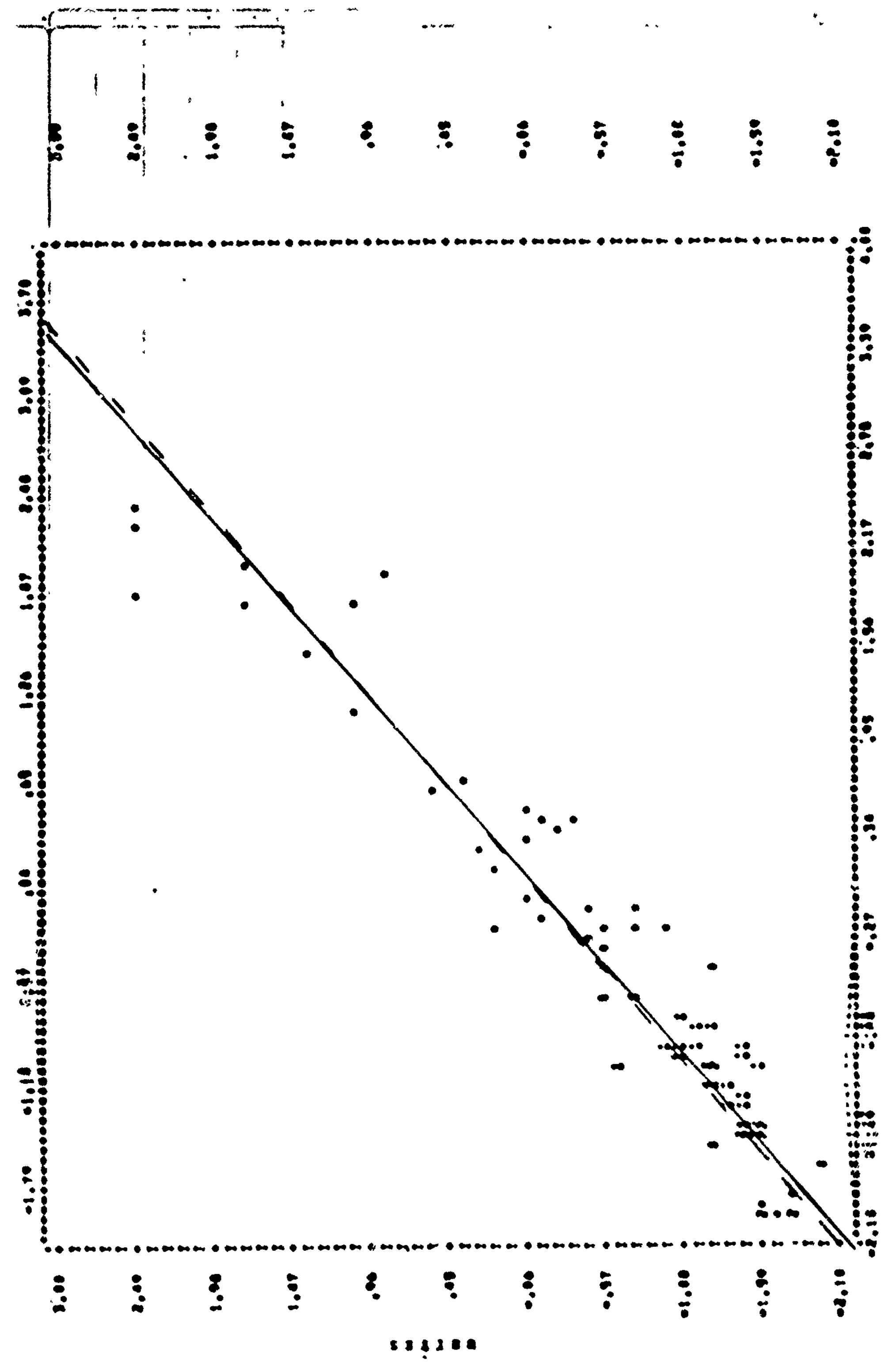


FIGURE 6. SCATTERPLOT OF THE ITEM b PARAMETERS OF BLACKS AND WHITES.

S/N 0190-LF-003-6100
3IND-GEN 1551/1 (12-60)

CELL *Ns* and r_ϕs FOR A TWO-BY-TWO TABLE CATEGORIZATION OF RGL AND RECRUIT ATTRITION CHARACTERISTICS OF BLACKS AND WHITES

Racial Group	Categories in Two-by-Two Table				r_ϕ	X^2
	<4.0 Attrition	<4.0 Non Attrition	≥4.0 Attrition	≥4.0 Non Attrition		
FR Data Set						
Blacks	4	16	40	891	.11**	10.94
Whites	13	8	60	912	.31***	95.43
TR Data Set						
Blacks	4	15	31	630	.12**	9.79
Whites	15	8	109	601	.23***	38.78

** *P* <.01
*** *P* <.001

PERCENTAGE OF REFINED SAMPLE OMITTING ITEMS

No. Items Omitted	FR DATA SET		TR DATA SET	
	Black	White	Black	White
Part 1 (Vocabulary)				
5	5	1	6	5
10	2	0	4	3
15	1	0	2	0
20	1	0	2	0
25	1	0	2	0
Part 2 (Reading Comprehension)				
5	7	1	11	6
10	2	0	4	2
15	1	0	2	1
20	0	0	0	0

PERCENTAGE OF COMPLETE SAMPLE OMITTING ITEMS

NO. ITEMS OMITTED	FR DATA SET BL.	FR DATA SET WT.	TR DATA SET BL.	TR DATA SET WT.
Part I (Vocabulary)				
5	9	4	12	8
10	5	2	7	5
15	2	1	3	2
20	1	0	2	1
25	1	0	1	1
Part 2 (Reading Comprehension)				
5	13	5	17	10
10	7	3	9	7
15	3	1	4	3
20	1	1	2	2
25	0	1	1	0

Categorization of Blacks and Whites in the TR Data Set

In Terms of $\hat{\theta}$ Ability Estimates

	Discharged--Recruit Training				Not Discharged--Recruit Training			
	Black		White		Black		White	
RGL	$\hat{\theta}$	M.D.	$\hat{\theta}$	M.D.	$\hat{\theta}$	M.D.	$\hat{\theta}$	M.D.
2				6	1	1		3
3		4		9	2	11		5
4		3		11	11	28		14
5	5		5	8	73	4	18	20
6	4	1	10	3	82		45	2
7	4		9		96		65	
8	7		13		82		106	
9	4		12		103		122	
10	3		17		114		141	
11			21		37		64	
12							4	
Total	35		112		645		609	

INITIAL TEST OF ARMORED CAVALRY ENGAGEMENT SIMULATION

Claramae S. Knerr
US Army Research Institute for the Behavioral
and Social Sciences

MAJ Angelo A. Severino
US Army Training Support Center

John J. Bosley
Human Sciences Research, Inc.

Engagement simulation is a generic term for training techniques that provide realistic tactical training under conditions that simulate the complex modern battlefield. The emphasis for the realism is on the psychological fidelity of the training environment and procedures (Root, 1976). Fidelity factors include the cues to which the soldiers must respond, their opportunity to respond, and changes in the situation as a result of their actions. Three characteristics of engagement simulation (ES) exercises contribute to psychological fidelity: (a) they are two-sided, free-play tactical exercises, with (b) objective, realtime casualty assessment, and (c) simulation of all weapons effects and signatures.

The earliest type of ES, called SCOPES (for Squad Combat Operations Exercise, Simulated), was developed for infantry squads. In SCOPES exercises, squads conduct two-sided free-play exercises, so that each force opposes a motivated, intelligent enemy. Objective casualty assessment is achieved when a soldier, looking through a six power telescope mounted on his M16 rifle, correctly reads a three inch, two digit number on the helmet of an opposing unit member. The power of the telescope and the size of the helmet number are calibrated to produce

hit/kill probabilities realistic for the weapon's lethality. A casualty is assessed when the soldier fires a blank round and correctly identifies the opposing helmet number. The soldier must fire a blank. If not, it is considered a misfire and no casualty is assessed. A controller with the fire team radios the helmet number to the controller with the opposing element, who informs the target soldier. Soldiers who have been "hit" must remove their helmets, lie down, and not communicate or otherwise participate in the exercise.

Physical fidelity of this casualty assessment method cannot be considered high. The only indication that soldiers have that they are casualties is when a controller tells them that they have been hit. However, because they know that casualties are assessed using strict rules, they know that they have performed incorrectly (e.g., did not stay under cover). Therefore, the situation has psychological fidelity. Soldiers learn very quickly to low crawl.

Procedures to conduct tactical ES exercises with combined arms elements have been developed and implemented under the name REALTRAIN. Procedures for objective casualty assessment have been established for the M60 machinegun, hand grenade, M18A1 Claymore, M16A1 anti-personnel and M21 anti-tank mines, tank main gun, and light, medium, and heavy anti-tank weapons (LAW, DRAGON, and TOW). For weapons with longer range than the M16 rifle, the controller is equipped with optics to sight the individual helmet numbers, or numbers on panels attached to the vehicles. For example, tank controllers have ten power breech mounted telescopes, and controllers with TOW gunners have ten power telescopes mounted on the TOW sight.

Indirect fire, either mortar or artillery, is simulated by detonating artillery burst simulators at the actual impact location requested by the players. The artillery simulators are delivered by fire markers, usually mounted in jeeps. Controllers with the players assess casualties within the "kill radius" of the simulated rounds when the simulators are detonated. For example, exposed soldiers within a 50 meter radius of a simulated 4.2 inch (107 mm) mortar burst are assessed as casualties, and vehicles lose communications although they are not destroyed. The communication loss enhances the psychological fidelity by simulating confusion caused by indirect fire.

The sights and sounds of battle are represented by pyrotechnics. Each soldier's weapon, crew-served weapon, and armored vehicle is equipped with pyrotechnics to simulate the flash and noise of the weapon signature. These signature simulators provide an important aspect of the psychological fidelity. The firer may have an excellent position, but upon firing the signature changes the situation by cuing the enemy as to the firer's location. The firer is likely to be "hit" unless he moves after firing. The signature simulators do not exactly reproduce the actual weapon signature, however they do produce situation changes that necessitate realistic player responses.

ES training entails three stages: (a) a free-play tactical exercise, (b) After Action Review (AAR), and (c) successive repetitions of the exercises and AAR. The exercise provides performance training under realistic tactical conditions in a discovery, or trial and error, paradigm coupled with structured feedback. Each exercise is followed by

an AAR which recreates the action and provides additional information to the soldiers as to the consequences of their actions. Soldiers who "killed" another soldier or vehicle describe how they detected and "destroyed" the enemy. Soldiers who were "casualties" hear from their peers the errors that led to their being "hit." Although disagreements arise, which are sometimes very spirited since motivation and competition are high, the objective casualty assessment system indicates convincingly "killers" and the "killed." Feedback from the opposing force reinforces learning through peer dialogue. The AAR leader guides the discussion, but does not critique or lecture.

The AAR leader is usually a senior controller, who is not assigned to a participating vehicle, but serves to coordinate the controllers, control the exercise as a whole, and act as the unit commander. The AAR leader uses a record of the casualties to guide the discussion, which recaps the exercise chronologically. The record is maintained in an exercise control station where personnel write the time and elements "hit" as the controllers report them on the radio. For individual soldier casualties, for example, the controller reports "29 killed by 45, 29 killed by 45." The controller with individual 29 acknowledges the "hit" by reporting "29 confirmed, 29 confirmed." The exercise net control station (NCS) recorder writes the time, target number, firer number, and checks that the "hit" was confirmed. Thus, the sequence of casualties is recorded for the AAR leader to guide the discussion.

Between the exercise and the AAR, the AAR leader meets with the controllers to review the NCS record, correcting and augmenting it to enhance the AAR. This controller debrief is used to settle controversies over "hits" and derive training points to emphasize in the AAR.

Validations of both the infantry squad SCOPES and combined arms REALTRAIN indicate that ES training is effective in achieving tactical proficiency. Tactical ES trained units improved in aspects of tactical proficiency such as (a) maximizing effects of available weapons on the enemy, (b) minimizing effects of enemy weapons, (c) effective intra- and inter-unit coordination, and (d) adaptive response to enemy actions in a dynamic combat situation. In the SCOPES validation, performed this May at Fort Ord, SCOPES trained squads were compared with conventionally trained squads (Banks, J.H., Hardy, G.D., Scott, T.D., Kress, G., and Word, L.E., 1977). In the REALTRAIN validation, performed in Europe in 1975-1976, combined arms units with three weeks of ES training were compared with similar units in their first week of ES training (Root, R.T., Epstein, K.I., Steinheiser, F.H., Hayes, J.F., Wood, S.E., Sulzen, R.H., Burgess, G.G., Mirabella, A., Erwin, D.E., and Johnson, E., 1976). In addition to the performance indicators listed above, the controllers and participants reported that, in their opinions, the ES exercises provided effective training (more effective than conventional training).

The nature of armored cavalry presented a threefold challenge for ES development: the reconnaissance function, a combined arms composition, and the inclusion of mortar. First, the armored cavalry functions as the "eyes and ears" of the maneuver forces, performing reconnaissance

missions, i.e., information gathering and reporting. These missions do not lead to the casualty producing engagements typical of other maneuver arms tactical training. In some instances, they are one-sided, with no firing or casualty assessment, thus no weapons effects or signature simulators. Thus, all three aspects of ES that enhance psychological fidelity would be inoperative. An example would be reconnaissance of an area that does not contain enemy elements. In other instances, enemy elements may be present, so that the exercise is two-sided. If the opposing elements fire, then the exercise converts to a casualty-producing ES training, and the standard ES procedures apply. On the other hand, if they do not fire, but continue to perform information gathering and reporting functions (e.g., reports of enemy detection) then the reconnaissance activities can be reenacted in the AAR. However, without special techniques, the AAR dialogue would be the opinion of one opposing force against the other, still lacking the objective assessment that makes ES casualties credible and convincing.

Second, the armored cavalry platoon is the smallest combined arms force in the Army, containing scout, light armor, infantry, and mortar sections. The task of simulating the entire array of armored cavalry weapons set the scope of the development, with the complexity of the exercises a major concern. For successful development and eventual implementation, the cavalry ES system had to be as simple as possible for units to employ.

Third, the mortar section is organic to the armored cavalry platoon, therefore it was included in the tactical exercises. In past ES exercises indirect fire elements were merely simulated, and fire markers

delivered artillery burst simulators to indirect fire impact locations. In contrast to the previous indirect fire methods, the mortar section was physically present with the maneuver forces in the armored cavalry exercises.

In addition to the aspects unique to armored cavalry, all of the usual aspects of engagement simulation, described below, required development.

Weapon effects and signature simulation. New hardware and accompanying procedures for its use were developed for some weapons, including the M551 Sheridan, Armored Reconnaissance Airborne Assault Vehicle (with conventional HEAT and Shillelagh missile), M114 scout vehicle, Armored Command and Reconnaissance Carrier (with 20mm cannon), and the 4.2 inch (107mm) mortar on the M106 Armored Mortar Carrier. Signature simulators, controller optics, and rules for their use were devised.

Exercise control. Controller duties, rules of engagement, casualty assessment, and controller communications were tailored to the vehicle type, crew, and weapon system. Each vehicle had one controller, except for the infantry M113 which had two controllers, one for each fire team. Each opposing force had a senior controller who functioned as the next higher unit commander. The senior controller acts as the troop commander when armored cavalry platoons are the opposing forces.

Exercise recording. In typical REALTRAIN exercises, Net Control Station (NCS) personnel record the simulated casualties and confirmations. To incorporate reconnaissance information, sightings (any detection of enemy activity and elements) were also reported, confirmed,

and recorded on the NCS record. Other methods for recording reconnaissance information were field notes, kept by the vehicle controllers, and logs of the tactical radio nets.

After Action Review. The REALTRAIN NCS records, tactical notes, and reconnaissance information were compiled during the controller debrief after the exercise, and used as input to the AAR. More emphasis was placed on information gathering and reporting than in AARs for typical REALTRAIN exercises.

OBJECTIVES

The overall objective was to develop engagement simulation for armored cavalry. Specifically, the objectives in testing the candidate procedures were to examine:

1. Procedures designed to emphasize the reconnaissance functions in ES exercises,

2. Procedures for incorporating reconnaissance functions into the controller debrief and the After Action Review,

3. Controller procedures and the control system, and

4. Effectiveness of the weapons effects and signature simulators for armored cavalry weapons.

METHOD

To meet the research objectives, the following types of instruments, described in the paragraphs below, were developed for data collection:

1. Records of information gathering, and reporting functions, and

2. Indirect measures such as attitudinal data concerning the procedures, simulators, training value, and AAR.

Records of information gathering and reporting functions. A variety of procedures were tested to incorporate reconnaissance functions into the exercises. Some of the forms used to gather data on these functions are forms primarily used in support of the ES training method, for recording the exercise events. The first one was the casualty record sheet typically maintained by the net control station (NCS) during the exercises. The NCS record includes the target, firer, time, and confirmation of each casualty. This casualty record sheet was altered to include reports of enemy detections (e.g., by sighting the enemy) in addition to casualties. The detection was called over the exercise control net, analogous to the call of a casualty. The target (sighted enemy element, in the case of sightings), firer (element that sighted the enemy), time, and confirmation were recorded on the NCS sheet. The altered sheet is shown in Figure 1.

The senior controllers kept notes during the exercises, largely critical incidents and reconnaissance information from the troop tactical nets. The senior controller who conducted the AAR used these notes to reconstruct the action and to focus discussions of the reconnaissance functions.

Printed three by five cards were prepared for the vehicle and infantry fire team controllers (Figure 2). The controllers were instructed to write the individual helmet and vehicle numbers on one side

CONTROLLER CASUALTY/SIGHTING RECORD SHEET

Day ____________ Period ____________ Exercise No.____________

Senior Controller: Green Force ____________ Brown Force ____________

Senior Controller ____________ ____________ Ops NCO ____________
(Name) (Name)

GREEN FORCE				BROWN FORCE			
Vehicle Nos.	Helmet Numbers			Vehicle Nos.	Helmet Numbers		

(The above portion will be filled out *before* the exercise.)

EVENT SEQUENCE RECORD

Target or Observed Element	Firer / Observer	Time of Fire, Sight	Confirm	Comments, Notes

Fig. 1. Revised NCS Record

Front of 3 x 5 Card

Day ________ Period ________ Ex. No. ________

Vehicle No. ________ Green ________ Brown ________

Rank/Name ________________________________

GREEN Vehicle Numbers:	BROWN Vehicle Numbers:
Helmet Numbers:	Helmet Numbers:

Back of 3 x 5 Card

Frag. Order:

Troop Questions:

Material Equipment Problems:

Locations (Sightings & Where Seen):

Fig. 2. Controller Field Note Card

of the card at the start of the exercise. This form was designed to encourage the controllers to keep records of the numbers, particularly those they control, and the casualties. These practices had proven useful in previous ES exercises. On the reverse side of the card, controllers were instructed to write notes to assist their participation in the controller debrief (orders, plans, troop questions, problems, and enemy detections).

Indirect Measures. Questionnaires were developed to record subjective judgments of the participants and controllers. Participants were asked to rate training value, simulator credibility, and utility of the candidate procedures. Vehicle and infantry fire team controllers were asked about casualty assessment and other ES procedures, hardware utility, simulator credibility, controller debrief and AAR, and training value of the exercises for the controllers.

PRELIMINARY FIELD TESTS

Procedures were drafted for armored cavalry ES and were examined and revised in a series of field tests. These field tests were developmental in nature, rather than validations of a completed system. Although data were collected whenever possible, no formal experiments were conducted. Field validation, as conducted for SCOPES and REALTRAIN awaits completion of an initial system for armored cavalry ES.

Figure 3 summarizes the field tests for Armored Cavalry ES. Some small scale exploratory tests were run to examine the draft procedures and the hardware devised to simulate the armored cavalry weapons. The

HOST UNIT	LOCATION	MONTHS (1977)	VEHICLES TESTED	OTHER OBJECTIVES
BNCOC	FT HOOD	JAN-FEB	SCOUT VEHICLE M151, M60 MG	PRETEST RECON PROCEDURES
BNCOC	FT BLISS	APR	SHERIDAN M551	REFINE DATA INSTRUMENTS PRETEST NEW RECON PROCEDURES AAR EMPHASIS ON RECON
3D ACR	FT BLISS	MAY	SCOUT VEHICLES M114, 20MM CANNON M113, TOW SHERIDAN, M551 MORTAR, 4.2 IN (107MM)	RECON DATA RECON AAR EXERCISE RECORDING EXERCISE CONTROL
3D ACR	FT BLISS	NOV		FOCUS ON RECON

FIGURE 3. DEVELOPMENTAL TEST SUMMARY

exploratory tests were done during Basic NCO Courses (BNCOC) at Ft Hood and Ft Bliss. BNCOC prepare soldiers for squad leader (E6) positions in infantry, armor, artillery, combat engineer, and air defense. The course contains three days of ES exercises.

Ft Hood BNCOC Exercises. The Ft Hood BNCOC ES exercises contained infantry squads and tanks, with approximately ten vehicles per exercise. The instructors added a scout squad to one of the opposing forces and used it with no difficulty. The M60 machineguns mounted on the scout vehicles had been simulated in engagement simulation before, and the procedures were satisfactory for the scouts. These exercises were an important first step in the armored cavalry engagement simulation development. Although no formal data were collected, the inclusion of scouts was obviously feasible. The BNCOC instructors provided ideas that were very valuable for further development of armored cavalry ES and they tried out data collection forms prior to use in larger exercises.

Ft Bliss BNCOC exercises. At the Ft Bliss BNCOC, tests focused on integration of the reconnaissance functions and initial use of procedures for the Sheridan M551 weapons effects and sig re simulators. Three exercises were conducted employing scouts, light armor and infantry squads with approximately eight vehicles per exercise. BNCOC students served as controllers, and some of the instructors assisted with the managing of the exercises. Their highly skilled assistance in small, easily manageable exercises facilitated the examination of new procedures.

The scouts were mounted in M113s with .50 caliber machineguns and in the same type of vehicles that had been tested at Ft Hood. Since both of these vehicles and their weapons had been tested in ES exercises before, it was simply verified that their simulation was satisfactory. The primary vehicle to be examined during the Ft Bliss BNCOC exercises was the Sheridan M551.

The intended signature simulator for the Sheridan main gun is the Hoffman device, which has been used successfully as the M60 tank main gun simulator in previous ES exercises. Unfortunately, Hoffman rounds were not available for these tests. The substitute was an M116 hand grenade simulator, detonated to simulate the noise and flash of the gun. The Hoffman device, which provides more realistic noise and flash, is the preferred simulator.

A modified missile aft cap, with a ten power telescope inserted in the center, was used in the breech of the main gun as the controller telescope. It seemed to be satisfactory during the BNCOC exercises.

Some methods for incorporating the reconnaissance functions were pretested at the Ft Bliss BNCOC. Enemy detection information was reported over the exercise control net. A large number of sightings (25) were reported but few were confirmed (4). These reports contributed little to the AAR. Reporting them over the exercise control net substantially increased the load on the net depending on the number of reports attempted.

Vehicle and infantry fire team controllers in the exercises were provided with 3x5 cards for them to record the ES numbers and their notes for the controller debrief. Almost all of the BNCOC controllers

used the cards. The controllers used the cards in 19 of the 20 possible instances: 18 of these were used for notes for the controller debrief, and 11 were used to record the ES numbers used during the exercises.

The senior controllers kept notes during the exercise or had assistants in their jeeps to help them keep the notes. They used these notes during the AARs to reconstruct the action. They incorporated the notes with the maneuver exercise control net records which contained sightings reported over the control net. However, even with all these sources of data, or perhaps because of it, it was very difficult to incorporate the reconnaissance information into the AAR. There were too many sources of information and the functions were too complex to bring together quickly after the exercise in the field environment.

The BNCOC participants and controllers completed questionnaires (described in the Indirect Measures section of this report) and commented on them. Their responses and comments were used to revise the questionnaires prior to their use in larger exercises.

PLATOON EXERCISES

Armored cavalry ES procedures were revised on the basis of the BNCOC results, and tested in May, 1977, with troop support provided by the 3d Armored Cavalry Regiment (ACR), Fort Bliss, Texas. C Troop, 1st Squadron, was the test unit.

Controller Training. The first three days of the two week test were devoted to controller training, in which C Troop personnel were trained in ES procedures, controller duties, and the After Action

Review. Approximately eight hours of practical exercises were conducted for controllers and participants to practice their duties. In these practical exercises, the platoon was divided into sections, so that the opposing forces were scouts versus scouts, infantry versus infantry, and Sheridans versus Sheridans.

A controller communication exercise was conducted to familiarize the controllers with ES procedures and duties. The full complement of controllers practiced sending, receiving, and confirming typical ES control messages over a radio net prior to the first full-sized exercise they controlled. The transmissions were tape recorded and played back for discussion. The session, with play back, was conducted twice. This training was evaluated so favorably that it was incorporated forthwith in the REALTRAIN implementation program.

Exercises. Six days of platoon versus platoon exercises were conducted, with one exercise and After Action Review each day. The armored cavalry platoon composition, with vehicles organic to the test unit are shown in Figure 4. Due to staffing levels and maintenance requirements, fewer than the full complement of ten vehicles per platoon participated in some of the exercises. Each platoon in C Troop participated as one of the opposing forces in four exercises (Table 1), and as controllers in the other two exercises.

Missions. Missions were selected from the Army Training and Evaluation Program for Armored Cavalry Squadron and Armored Cavalry Troop (ARTEP 17-55), with assistance of 1st Squadron personnel. The

HEADQUARTERS SECTION

PLATOON LEADER
M114
(20MM CANNON)

SCOUT SECTION

SCOUT SQUAD	SCOUT SQUAD
M114	M114
(20MM CANNON)	(20MM CANNON)
M113	M113
(TOW)	(TOW)

LIGHT ARMOR SECTION

3 - SHERIDANS, M551
(152MM CONVENTIONAL ROUND AND MISSILE)

RIFLE SQUAD
M113
(.50 CAL. MACHINE GUN)

MORTAR SQUAD
M106
(4.2 INCH (107MM) MORTAR)

FIG. 4. ARMORED CAVALRY PLATOON COMPOSITION DURING ES EXERCISES.

TABLE 1. PLATOONS AND MISSIONS BY EXERCISE

EXERCISE	PLATOON	MISSION	PLATOON	MISSION
1	1	ROUTE RECON	3	SCREEN
2	1	DELAY	3	ZONE RECON
3	3	ROUTE RECON	2	SCREEN
4	3	DELAY	2	ROUTE RECON
5	2	ZONE RECON	1	DELAY
6	2	ROUTE RECON	1	SCREEN

missions were those appropriate for a regimental squadron, were considered to be of training benefit to C Troop, and emphasized the reconnaissance functions. They are representative of level 1 ARTEP missions, i.e., those comprising the minimum acceptable performance for a combat-ready, full strength unit. Missions were paired in each exercise so that one platoon had a reconnaissance mission while the opposing platoon had a screen or delay mission (Table 1). The platoon with the screen or delay mission had time to prepare a position before the opposing platoon moved to contact. The 1:1 force ratio was tactically unrealistic for what amounted to an attack against a prepared defense, but it was highly desirable to train each platoon as a unit.

Terrain. The training area was flat desert, having only 40 feet difference between the high and low elevation. It was dotted with sand dunes and low scrub vegetation. Unpaved trails were the only features that assisted in position location and they were visible for only short distances because of the sand dunes. The exercise lanes were approximately three by six kilometers. The major axis of each exercise lane followed one of the trails. Position location proved to be very difficult and unreliable on this terrain.

RESULTS AND DISCUSSION

Target reports and confirmation. The objective casualty system is a primary strength of ES training, as described in the introduction. Confirmed casualty information, with certainty as to who engaged whom,

provides immediate and definite feedback. In contrast to typical ES exercises, where virtually all of the casualties are reported by number and confirmed, only a third (31 of 104 targets) of the casualties were reported by number and confirmed in this test (Table 2). Of the 104 targets reported during the six exercises, 38% were reported by ES number, while 62% were reported by coordinates.

Terrain characteristics appeared to be responsible for the low number of target reports by number. The opposing forces were unable to maneuver without being detected (due to vehicle exhaust smoke or dust clouds) at long ranges. Thus, vehicles were engaged either at ranges beyond those in which the numbers were legible, or when sand dunes obscured the number panels. The controllers cited problems in identifying the ES numbers of opposing vehicles, giving the engagement distances as the main reason. Of the 48 controllers who responded to the question, 29 (60%) reported that the enemy vehicles were too far away to read the number.

Only 58% of the targets were confirmed (60 of the 104 targets reported). The percent of confirmations was significantly higher for targets that were reported by ES number (78%) than were reported by coordinates (45%: $z = 3.18$, $p < .01$). Targets are easier to confirm when they are reported by ES number, since the controller on the specified opposing vehicle can hear and respond to the radio message. Confirmation of a target reported by coordinates requires that the senior controllers carefully check vehicle positions, and contact the possible

TABLE 2

Target Reports and Confirmations

Confirmation	Targets Reported		
	By ES Number	By Coordinates	Total
Yes	31	29	60
No	9	35	44
Total	40	64	104

target vehicles individually in an attempt to confirm the casualty. Lack of terrain features, and inexperience of the controllers, made the location determinations difficult and inaccurate. The senior controllers had to locate the vehicles, often by extensive radio use, increasing the load on the exercise control net substantially over the load in typical exercises. These additional transmissions taxed the senior controllers, who were responsible for troop command as well as exercise control. Transmission load on the control net degrades the exercises by interfering with controller reports and confirmations of casualties. Slow or inaccurate removal of elements reported as targets decreases the realism during the exercise, and makes reconstruction of the action in the AAR less convincing. For example, since accurate coordinates were difficult to determine, the crews of target vehicles were not convinced that their vehicles were the ones reported as targets, especially if other vehicles were nearby. Thus, reinforcement value from the objective, definite casualty system was decreased in approximately half of the simulated engagements during these exercises.

Results of incorporating reconnaissance functions. Enemy detection information was reported over the exercise control net in the first two platoon exercises. Only 4 sightings were reported, and only one of these was confirmed. These reports contributed little to the AAR, while they increased the load on the control net. Due to the low usefulness and interference with the control net, sighting reports were discontinued after the second platoon exercise.

The AAR leaders found the notes that they kept during the exercises to be the most helpful tactical record during the controller debrief and the AAR. Such notes are difficult for the senior controllers to maintain, since they are traveling over rough terrain, and since they have the additional responsibility of functioning as the unit commander. Future emphasis will be on improving methods and use of the senior controller notes.

Vehicle and infantry fire team controllers in the platoon exercises were provided 3 by 5 cards to record ES numbers and notes for the controller debrief. The controllers used almost all of the cards (111 of 120, or 93%). They used over half of the cards to record the ES numbers (68%), and just less than half to record notes for the controller debrief (42%). About a third of the cards had both ES numbers and notes (37%). This field note card usage is high compared to usual paper work in field exercises.

The high usage rate is corroborated by the controllers' ratings (N=48) of the field note card utility (including use for both ES numbers and notes):

Good	58%
Fair	33%
Poor	8%

Overall, the controllers (N=48) reported favorably on the usefulness of the field notes:

Very helpful	44%
Somewhat helpful	19%
Not helpful	6%
Didn't take notes	25%
No response	6%

During the six platoon exercises, 91 tactical reports were recorded from the troop tactical radio net. Approximately half (45) were reports of enemy sightings, which represent an average of 7.5 reports per exercise. The records proved too voluminous for the AAR leader to organize prior to, or during, the controller debrief. However, procedures are being drafted to test in the next field experiment to enable the AAR leader to use the tactical radio net records in order to reconstruct the action, especially as needed for the reconnaissance functions.

Casualty Assessment. Casualty assessment rules, printed on cards, were distributed to the controllers for their use during the exercises. These cards, used to reinforce the casualty assessment training, appeared to be effective. The controllers reported that they had no problems with casualty assessment (39 of the 48 controllers who answered the question, or 81%, marked the response category "no problems"). Their reports were consistent with observations by the training advisors and research personnel.

Vehicle Casualties by Mission. Table 3 presents vehicle casualties by mission type. Platoons assigned reconnaissance missions (zone or route) lost 65% of their vehicles, while platoons with screen or delay

Table 3

Simulated Engagements by Mission

Mission	Nr. Vehicles Played	Nr. Vehicles "Hit"	% Vehicles "Hit"
Route Recon.	36	22	61%
Zone Recon.	18	13	72%
Recon. Total	54	35	65%
Screen	26	11	42%
Delay	27	7	26%
Prepared Position Total	53	18	34%

missions lost only 34%. These outcomes appear realistic given the 1:1 force ratios of the moving and defending elements. When equal forces meet in battle, the moving force is expected to be at a disadvantage, compared to the force in a prepared position. The realistic outcome statistics attested to the realism of the exercise itself.

Weapons Effects and Signature Simulators. Procedures and hardware for simulating the M551 Sheridan main gun, M114 scout vehicle 20mm cannon, and 4.2 inch (107mm) mortar were evaluated. The M113 armored personnel carrier with either the .50 caliber machinegun or TOW were also played, but their evaluation was not a primary issue because they have been played in past ES exercises.

Table 4 shows that the TOW missile inflicted the largest number of vehicle casualties: 23 of the total 53 vehicles "destroyed." The TOW missile has long range and high lethality, and so accounts for a high portion of the casualties. The TOW scout vehicle is among the leading elements of the platoon, therefore, it contacts the opposing force early in the exercise. In this test, both the TOW missile and .50 caliber machinegun mounted on the same vehicle were used effectively.

M114 Scout vehicle with M139 20mm gun/cannon. The 20mm cannon signature was simulated by an M117 flash simulator. Several of the simulators were attached to a board on the M114 scout vehicle front, and were detonated by pulling a trip wire. The simulator was easy to hear, but did not ideally represent the gun signature. Safety was a major problem as noted in incidents such as accidental firings. The M117 is an interim device to be used only until a signature simulator is developed for the 20mm cannon.

TABLE 4

SIMULATED ENGAGEMENTS BY WEAPON TYPE

	Platoon Leader M114		Scout M114		Scout TOW M113		Sheridan M551		Mortar M106		Infantry M113		Total	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Vehicle Targets	6	50	14	67	13	54	14	48	2	17	4	44	53	50
Vehicles Played	12		21		24		29		12		9		107	
Simulated Engagements By:														
M114 20mm Gun			1	7	2	15	2	14					5	9
TOW Missile	2	33	7	50	5	38	7	50			2	50	23	43
.50 Cal Machinegun on TOW M113	2	33	1	7	2	15	1	7					6	11
Sheridan 152mm Gun			2	14			1	7					3	6
Mortar	2	33	3	21	2	15	1	7	2	100	1	25	11	21
LAW					2	15	2	14			1	25	5	9

Controller optics for the 20mm cannon were fabricated from the TOW controller optics. A 10 power telescope was attached to the cannon above the gunner's 13 power sight. During the two weeks of the exercises, threads in the mounting block became damaged so that the telescope worked loose and did not remain aligned with the gunner's sights. Thus, the controller had a different sight picture than the gunner, and could not identify targets properly. The mount is being improved to solve this problem.

M551 Sheridan with 152mm gun/missile launcher. The same signature simulator (M116 hand grenade simulator) and controller optics were used as described for the BNCOC tests. The controllers and participants reported favorably on the hand grenade simulator (e.g., easy to hear, realistic simulation of the main gun), but they also suggested that the signature simulation be improved as to loudness, flash, and smoke. When used in this armored cavalry application, the Hoffman device will provide these improvements and the necessary realism.

The modified missile aft cap, with a ten power telescope inserted in the center, proved unsatisfactory during the platoon exercises. During the exercises, the aft cap vibrated loose, and on occasions fell out of the breech. The missile aft cap has been further modified to correct this problem.

Sheridans contributed relatively little to the vehicle casualties (Table 4), despite the long range and high lethality of the main gun. They were held in reserve to react to enemy contact rather than joining

the casualty-producing engagements. This can be atrributed to the fact that the TOW section was well forward, and the tendency was to engage with the TOW because of its availability and to disregard the reconnaissance function.

<u>M106 Armored Mortar Carrier with 4.2 inch (107mm) mortar</u>. Two procedures were tested to incorporate the mortar section into ES exercises. One procedure used the M32 pneumatic training device, which attaches to the mortar and shoots a plastic round, via air pressure. The adjustment of air pressure determines the distance that the round travels, and it is set for each shot to represent the propellant charge. The direction of shot depends upon the sight deflection settings, just as it does a real round. A proportional conversion can be used to calculate, from the plastic round impact location, what would be the impact location of a real round. The procedure used in this test simplified the proportional conversion by laying out a rope scale with proportional distances marked for the range. Range was estimated by a controller, with the marked rope as a reference, and lateral distance (deflection) was estimated as distance from the rope to the right or left. The rope scale was difficult to use on the Ft Bliss terrain because the sand dunes interfered with placement.

In the alternate mortar procedure, a mortar controller observed the fire direction and gunnery procedures. When he detected errors, he computed an impact point and notified the fire marker to deliver the simulated rounds to the corrected location, rather than the location

requested by the observer. These procedures were too complex for one controller (observe two sets of crew members, compute impact points, operate the radio, and record the procedures). Future tests will examine assignment of two controllers, and simplified procedures.

The mortars, which remained in the rear of the armored cavalry platoons, were the vehicles least often engaged (Table 4). However, two mortars were hit by indirect fire from the opposing force. The first was a preplanned target, which the mortar had selected as its initial location. The training value regarding position selection was evident after this hit, since the mortar crew had selected a position that was a major terrain feature (trail junction) that was also a good point for an opposing force preplanned target. The crews discussed this issue, and quickly learned to select less obvious positions.

Fire Marker Transportation. Fire markers, who deliver the artillery burst simulators to the requested impact locations, usually travel in jeeps. In this test, an OH-58 helicopter was tried as the fire marker vehicle. Only one helicopter was employed (for safety over the small exercise lane), therefore only one of the opposing forces could have indirect fire simulation at one time. The helicopter had to leave the training area to refuel prior to the end of the exercises, terminating indirect fire support. The helicopter was on station approximately 90% of the exercise time.

The indirect fire simulation system produced eleven (11) simulated vehicle engagements, for 21% of the total hits. Note that mortar hits do not destroy vehicles, but knock out communications and kill exposed

personnel. Therefore, the 11 simulated vehicle engagements did not destroy the vehicles. In one case, a vehicle that had been hit by simulated mortar fire early in the exercise was destroyed by .50 caliber machinegun fire later in the exercise. Previous indirect fire simulation has shown a higher proportion of hits For example, during the REALTRAIN validation in Europe indirect fire accounted for 31% to 32% of the personnel and vehicle casualties. Various characteristics of the indirect fire simulation in these exercises at Ft Bliss appeared to reduce the mortar effectiveness. The problems in use of the helicopter were just described, and solutions will be tried in the next test. Other reasons for different indirect fire simulation effects include terrain, unit composition, and type of indirect fire simulated.

SUBJECTIVE TRAINING VALUE RESULTS

Participants and controllers were asked for their subjective evaluations of the training value of the ES exercises, and how the ES exercises compared with other training.

Participants (N=77) responded as follows to the question "How much would you say you learned during the training exercises you have just completed?":

A great deal	44%
Some	38%
Little or nothing	18%

When asked to compare the ES exercises to other training, most participants replied that the ES exercises were better:

REALTRAIN much better	36%
REALTRAIN better	43%
No difference	11%
REALTRAIN worse	11%

Compared to the REALTRAIN validation data from Europe, approximately the same portion of the questionnaire answers are in the combined "better" and "much better" categories, but in the Europe data, the majority responded that the ES training was "much more effective". Some differences in the responses may be due to scaling and administrative differences. It is possible that some of the training value, or at least the perception of the training value, was decreased by the problems that arose in conducting these armored cavalry exercises. Whether participants would report more perceived training value if the exercises were run better (e.g., improved target reporting and confirmation) remains to be tested in future exercises. It should be emphasized that the armored cavalry exercises entailed development of a new system, in contrast to the Europe validation of smoothly conducted training.

Controllers (N=48) were asked how much they learned about tactics when they served as controllers. Responses show that they perceive that they are learning, often as much or more than if they are part of the tactical team:

I certainly learned as much or more as a controller, as I would have if I'd been part of the tactical team.	54%
I learned a fair amount about tactics while acting as a controller.	33%
I didn't learn very much about tactics when I was controlling.	13%

Compared to the REALTRAIN validation in Europe, where 70% of the controllers reported that training value was much greater for controllers than for participants, the controllers were less positive concerning the training value of these exercises. These responses may reflect the exercise problems described above.

SUMMARY

This phase of testing was designed not to produce final answers but rather to explore and refine specific ES procedures for use by armored cavalry elements. While not emphasizing training effectiveness data at this point in the developmental sequence, perceptions of training value were collected from participants and controller personnel. Further, one could trace changes in tactical behavior over the series of exercises which would indicate that some learning had occurred. However, these measures do not represent the type of training effectiveness evaluation that would be conducted in a validation study. Performance measures appropriate for training effectiveness analysis will be tried in the next field test, but an objective training effectiveness analysis must wait for the validation.

These initial tests succeeded in determining several modifications necessary for the controller optics, signature simulators, and mortar controller procedures. The controller duties pertaining to casualty assessment appeared to be satisfactory. Given the modifications indicated, the casualty related aspects are ready to be written into the training program for armored cavalry ES.

All of the exercises in these tests contained a large number of simulated engagements. Thus, they were similar to typical ES exercises in that respect. However, a special emphasis must be placed on reconnaissance functions when dealing with armored cavalry. While the procedures for incorporating reconnaissance activities that were examined in the initial tests were a step in the right direction, additional development is required to fully play the reconnaissance functions. An approach containing several interrelated techniques is planned for the next field test. First, the exercise scenarios and operations orders will be designed to limit engagements and to foster reconnaissance behaviors. When the simulated engagements are limited, controllers can concentrate on observing and recording the information gathering and reporting activities of the elements that they control. The controller records, combined with records that appeared to be effective in the initial tests, are expected to increase objectivity. Without such records, the subjective and often conflicting judgments of the opposing forces constitute the only basis for discussion. Increasing the objectivity, or records of "ground truth" are expected to enhance credibility, and in turn increase the troop motivation and training value. Continued development of armored cavalry ES will focus on building the strengths of typical, casualty-producing ES into reconnaissance ES exercises. This revolves around realistic combat scenarios involving motivated opposing forces in an environment with strong psychological fidelity. Troops trained with ES may not have been in combat but they have had the opportunity to learn the lessons of combat without having to learn the hard way.

REFERENCES

Banks, J.H., Hardy, G.D., Scott, T.D., Kress, G., and Word, L.E. REALTRAIN Validation for Rifle Squads: Mission Accomplishment. ARI Research Report 1192. October 1977.

Root, R.T. Engagement Simulation: A Training Environment for the Learning of Complex Tactical Skills. Presented to the Military Operations Research Society, Ft Eustis, VA. December 1976.

Root, R.T., Epstein, K.I., Steinheiser, F.H., Hayes, J.F., Wood, S.E., Sulzen, R.H., Burgess, G.G., Mirabella, A., Erwin, D.E., Johnson, E. Initial Validation of REALTRAIN with Army Combat Units in Europe. ARI Research Report 1191. October 1976.

KNOWLEDGE TESTS OF MANUAL TASK PROCEDURES

William Osborn and Patrick Ford
Human Resources Research Organization

Paper For:

Military Testing Association Conference
San Antonio, Texas

October, 1977

KNOWLEDGE TESTS OF MANUAL TASK PROCEDURES[1]

The high cost of hands-on performance testing tends to complicate life for the developer of job proficiency tests. He is urged by reasons of economy to develop tests that are administratively feasible. This usually means tests that can be administered on a group basis--an interpretation that invariably leads to paper-and-pencil knowledge testing.

We know that knowledge tests are appropriate for tasks that are essentially mental, and we know they are inappropriate for tasks that involve finely tuned motor skill. But what of job tasks in between--tasks that involve both manual and mental activity? Many job tasks appear to be predominantly manual, but not particularly skilled. Placing some machines in operation, assembling objects, installing or repairing components, represent tasks that are essentially manual, but which, if performed without rigid time limits, cannot be considered psychomotor skills. This is not to say that such tasks require no skill. They must be learned, and if one identifies the skilledness of a task generally in terms of the amount of practice required to become proficient, then the aforementioned tasks are to some degree skilled. But the skilled aspect is probably mental, since knowledge must be acquired of what steps to perform, in what order and with what result. It may be hypothesized, in fact, that such manual task procedures can be performed with little or no practice, _if_ one _knows_ what, when and how to perform them.

If there is something to this hypothesis, proficiency can be measured validly in a knowledge testing mode, given one additional assumption: that the test medium is relatively neutral with respect to examinee differences in mental ability. This second assumption is necessary because we are considering a medium for testing that has no relevance to the medium for task performance. In other words, we would expect someone who can perform a task to be able to pass a hands-on test of that task; but if that person can't read or write at all well, we would be dubious of their ability to read and interpret written questions _about_ task performance. It seems important, therefore, when substituting for a hands-on test, that the substitute medium not favor one type of examinee over another. We should strive to use test media that are neutral with respect to task-irrelevant differences in abilities.

With this perspective, I would like to describe an experiment in which we evaluated the validity of knowledge tests as substitutes for hands-on tests of manual task procedures.

The experiment was designed to examine four methods of knowledge testing in terms of their relative and absolute correlation with hands-on task proficiency for high and low mental ability subjects (*Ss*). The specific research questions of interest were:

---

[1] This paper is based on research done under Contract No. DAHC 19-74-C-0059 with the U.S. Army Research Institute for the Behavioral and Social Sciences. Conclusions and opinions expressed are the authors', and not necessarily those of the U.S. Army.

- Do the four types of knowledge test correlate with hands-on task mastery?

- Do the types of test differ with respect to how well they distinguish masters from nonmasters?

- Do the types of test distinguish masters from nonmasters equally well for high and low mental ability levels?

- Do the types of test tend to produce the same kinds of errors in predicting task mastery?

Method

Test Development. Tests were developed for three Army tasks: Installation of the Field Telephone (TEL), Setting up a Mechanical Ambush with the Claymore (AMB), and Disassembling the M-16 Rifle (RIF). The first two are clearly low-skilled tasks. Rifle disassembly, however, would be classified more accurately as moderately skilled, since some of the steps entail manipulations that are not easily mastered in one or two trials. Each task was analyzed into steps on which the test items were based. In addition to a performance (hands-on) test, four versions of a knowledge test were developed for each task. One version was a conventional multiple-choice test. The other three employed pictures in an effort to minimize literacy demands, but used different methods of eliciting task knowledge. A description of the four tests follows.

- Written Choice (WC). This is a standard multiple-choice test consisting of one question for each step in the task. A question focused on recognition of how a step is performed, when it is performed, or what its correct outcome is. Alternative answers to a question were limited to realistic options; unrealistic distractors were avoided. The test was scored by giving one point for each correct answer; seven was the maximum possible score for the TEL and AMB tasks, and eight the maximum for RIF.

- Picture Choice (PC). This method included the same questions as the Written Choice, but photographs were used in place of the printed word in presenting answer alternatives. The possible points and scoring procedure were the same as for WC.

- *Picture Outcome (PO).* In this method a photograph of the result of an improperly performed task was presented. *Ss* were instructed to inspect the picture and circle any errors. This type of test focuses on recognition of correct task outcome only. Test score was based on one point for each error circled, minus one point for each non-error circled. Total score was not allowed to go below zero. The possible range of scores was from 0 to 4 for TEL and RIF, and 0 to 3 for AMB.

- *Picture Sort (PS).* Photographs of steps in task performance, including both correctly and incorrectly executed steps, were used in this test method. The pictures were scrambled and presented to *S* with instructions to select the correct steps and place them in the order they should be performed. This method was considered to be the most comprehensive in its coverage of task knowledge; what steps to perform, and how and when to perform them are required knowledge. The method relies on recognition, as do the others, but all task elements are tapped and the guessing factor is minimized. Scoring was based on the award of one point for each picture or group of pictures representing a correct step performed in proper sequence. If two correct steps were in improper order, credit was withheld for the first step. Steps were judged to be improperly sequenced only if it were impossible or hazardous to perform them in that order. Maximum possible score was seven for TEL, and eight for AMB and RIF.

Subjects. Thirty-seven soldiers from units at Fort Knox were tested. They were chiefly from combat arms MOSs and ranged in grade from E-2 to E-6. For the purpose of study design, *Ss* were in two mental ability (MA) groups: GT over 110 (high MA), and GT under 90 (low MA).[2] Twenty *Ss* were in the high MA group and 17 were in the low.

Procedure. On arrival at the test site the project was explained briefly to *Ss*. What was said to them took the following general form:

> We are working on a project to evaluate several different methods of testing. You will take a hands-on test for three tasks. Then you will take four other kinds of tests for each task. After the test we will ask your opinion of it. This is not an MOS test, so there

[2]The GT (General-Technical) is a combination of scores on a verbal and a quantitative aptitude test. It is considered to be the best indicator of general mental ability in the Army Classification Test Battery.

is no reason for you to be nervous. But the project is very important so, of course, we expect you to do as well as you can on every test.

All testing was done individually and began with administration of the hands-on test. At this point some *Ss* received training on the task before going on to the knowledge tests. This was done to control the range of task mastery within the two MA groups. The intention was to create a rectangular distribution of mastery, with approximately a third of each MA group being wholly unqualified on a task, a third being partially qualified, and a third full masters. This approach worked well at the full mastery level since only one *S* could perform a task (TEL) without further training. Thus, 7 masters were created in each MA group by training them to pass the three hands-on tests. The approach did not work as well within the nonmastery range since most *Ss* could perform some steps in the TEL and RIF tasks; only with the AMB task were any *Ss* trained to partial mastery.

Once an *S* had completed the hands-on test for a task, he was given the four knowledge tests successively. The order of test administration was counterbalanced over *Ss*

In addition to test performance, *Ss* were asked their opinions of the methods by having them rank them from 1 to 5 with respect to the question: "Do you think this test is a good way to find out if a soldier can (task statement)?"

Scores on the 15 tests--one hands-on and four knowledge tests for each of three tasks--and *Ss* ratings comprised the data that were analyzed.

Results

Continuous score correlations between knowledge test and hands-on performance for the three tasks are shown in Table 1 for the two levels of mental ability and for the total sample. With few exceptions the correlations are both statistically and practically significant. They are uniformly higher, regardless of test method, for the TEL and AMB tasks than for RIF, indicating that rifle disassembly is somehow different from the other tasks; a difference attributable perhaps to a more skilled motor component.

Comparison by type of knowledge test, for the total sample and and total performance on the three tasks, indicates that the Written Choice, Picture Choice and Picture Outcome correlate equally well (.83, .80, and .84 respectively) with hands-on performance. The Picture Sort method yields a somewhat smaller overall relationship (.58), although the reduction is attributable to the near-zero correlation for the RIF task. The trend toward higher correlations for total score than for task scores reflects a tendency for intercorrelations among tasks to be lower for a knowledge test than for the hands-on criterion.[3]

---

[3]The reader will recall that, by design, the same people were masters on all tasks (had maximum criterion scores) although nonmasters varied in degree of nonmastery from task to task.

TABLE 1

CORRELATIONS BETWEEN PERFORMANCE AND KNOWLEDGE TEST
METHOD FOR HIGH AND LOW MENTAL ABILITY GROUPS

MENTAL ABILITY GROUP	N		KNOWLEDGE TEST METHOD															
			WRITTEN CHOICE				PICTURE CHOICE				PICTURE OUTCOME				PICTURE SORT			
			TASK [a]															
			TEL	AMB	RIF	TOT	TEL	AMB	RIF	TOT	TEL	AMB	RIF	TOT	TEL	AMB	RIF	TOT
HIGH	20	r	.69	.55	.17	.79	.71	.66	.47	.82	.76	.77	.31	.78	.70	.62	-.31	.52
LOW	17	r	.73	.82	.75	.90	.80	.76	.51	.80	.68	.74	.65	.90	.79	.56	.29	.69
TOTAL	37	r	.71	.67	.49	.83	.75	.70	.51	.80	.72	.74	.55	.84	.72	.55	.04	.58

[a] TEL = Installing Field Telephone

AMB = Installing Mechanical Ambush with Claymore Mine

RIF = Disassembling M16 Rifle

TOT = Total Performance on the Three Tasks

Further analyses of the effectiveness of the different knowledge tests to distinguish masters from nonmasters, both within and between levels of mental ability, were carried out by analysis-of-variance. This is a reasonable way to examine the data, since mastery level was more of a manipulated "treatment" effect than a natural variate. Knowledge test performance, summed over tasks, of masters and nonmasters by mental ability level is shown in Table 2. All test methods did not have the same scale of measurement, so an ANOV (Winer, 1962) was performed on each method. Results of the four unweighted means ANOV are summarized in Table 3 and shown graphically in Figure 1. A clear and substantial main effect is revealed for mastery level, which merely represents the high correlations between knowledge test and task performance already mentioned. The size of this main effect for Picture Outcome relative to other test methods is worthy of note. The graphs in Figure 1 indicate that masters tend to average about five points higher than nonmasters on all tests, even though the <u>potential</u> range of performance on PO is only half that of the other tests. This would imply that a longer test would produce greater improvement in discrimination between masters and nonmasters for PO than for the other methods.

Performance on the knowledge tests tended to be lower for low mental ability *S*s than for high, as indicated by the slope of the curves in Figure 1. The difference is small, and in fact not statistically reliable according to the separate ANOVs. However, when performance was converted to standard scores within test method and aggregated over methods, the mental ability factor is marginally significant ($p > .05$). Moreover, the difference appears to be relatively constant over test methods (Figure 2), suggesting that no one method is superior in neutralizing mental ability differences.

One of the more interesting features of the data (Figure 1) is the trend, however slight, toward a larger difference between masters and nonmasters in the low MA group. This indicates a slightly higher correlation between knowledge test and task performance for low mental ability *S*s, a tendency also observed in Table 1 where for 9 of the 12 method/task combinations the correlation with mastery was higher within the low mental ability group. Note that this is not a statistically reliable phenomenon, but it suggests an interesting hypothesis: knowledge based tests predict task performance better among people of moderate to low mental ability than among those of high mental ability.

Validity in a strict correlational sense does not tell the whole story, however. The type of prediction or classification error is of practical interest. By converting knowledge test performance to pass-fail scores and arraying them against the master-nonmaster criterion, four-fold tables were generated from which the incidence of false negative (masters who failed the test) and false-positive (nonmasters who passed the test) classification errors were determined. The correlation and amount of classification error, of course, depend on the standard used in scoring pass-fail. Classification error was tabulated for a standard of full mastery on the knowledge test (pass = all items right) and again for a standard of part mastery

TABLE 2

KNOWLEDGE TEST PERFORMANCE (MEANS AND STANDARD DEVIATIONS)
OF MASTERS AND NONMASTERS BY TEST METHOD AND MENTAL ABILITY LEVEL

MASTERY LEVEL	MENTAL ABILITY		TEST METHOD WRITTEN CHOICE	PICTURE CHOICE	PICTURE OUTCOME	PICTURE SORT
MASTERS	HIGH	$\bar{X}$	18.71	20.28	10.29	18.71
		s	2.98	1.60	.76	3.25
		N	7	7	7	7
	LOW	$\bar{X}$	17.86	19.29	10.14	17.57
		s	1.95	2.10	1.21	3.41
		N	7	7	7	7
NONMASTERS	HIGH	$\bar{X}$	13.69	15.00	6.38	15.15
		s	2.56	2.80	1.61	4.56
		N	13	13	13	13
	LOW	$\bar{X}$	11.30	13.60	5.00	12.00
		s	1.83	2.59	1.41	3.06
		N	10	10	10	10

TABLE 3

ANOV SUMMARIES OF THE EFFECTS OF TASK MASTERY (*M*)
AND MENTAL ABILITY (*A*) ON KNOWLEDGE TEST PERFORMANCE

TEST METHOD	SOURCE	SS	df	MS	*F*
WRITTEN CHOICE	*M*	289.853	1	289.853	46.10**
	A	22.691	1	22.691	3.61
	M x A	5.126	1	5.126	.82
	Error	207.466	33	6.287	
PICTURE CHOICE	*M*	260.120	1	260.120	43.73**
	A	12.357	1	12.357	2.08
	M x A	.364	1	.364	.06
	Error	196.273	33	5.948	
PICTURE OUTCOME	*M*	177.034	1	177.034	95.38**
	A	5.060	1	5.060	2.73
	M x A	3.271	1	3.271	1.76
	Error	61.2483	33	1.856	
PICTURE SORT	*M*	180.178	1	180.178	12.73**
	A	39.781	1	39.781	2.81
	M x A	8.733	1	8.733	.62
	Error	466.939	33	14.150	

** $p < .01$

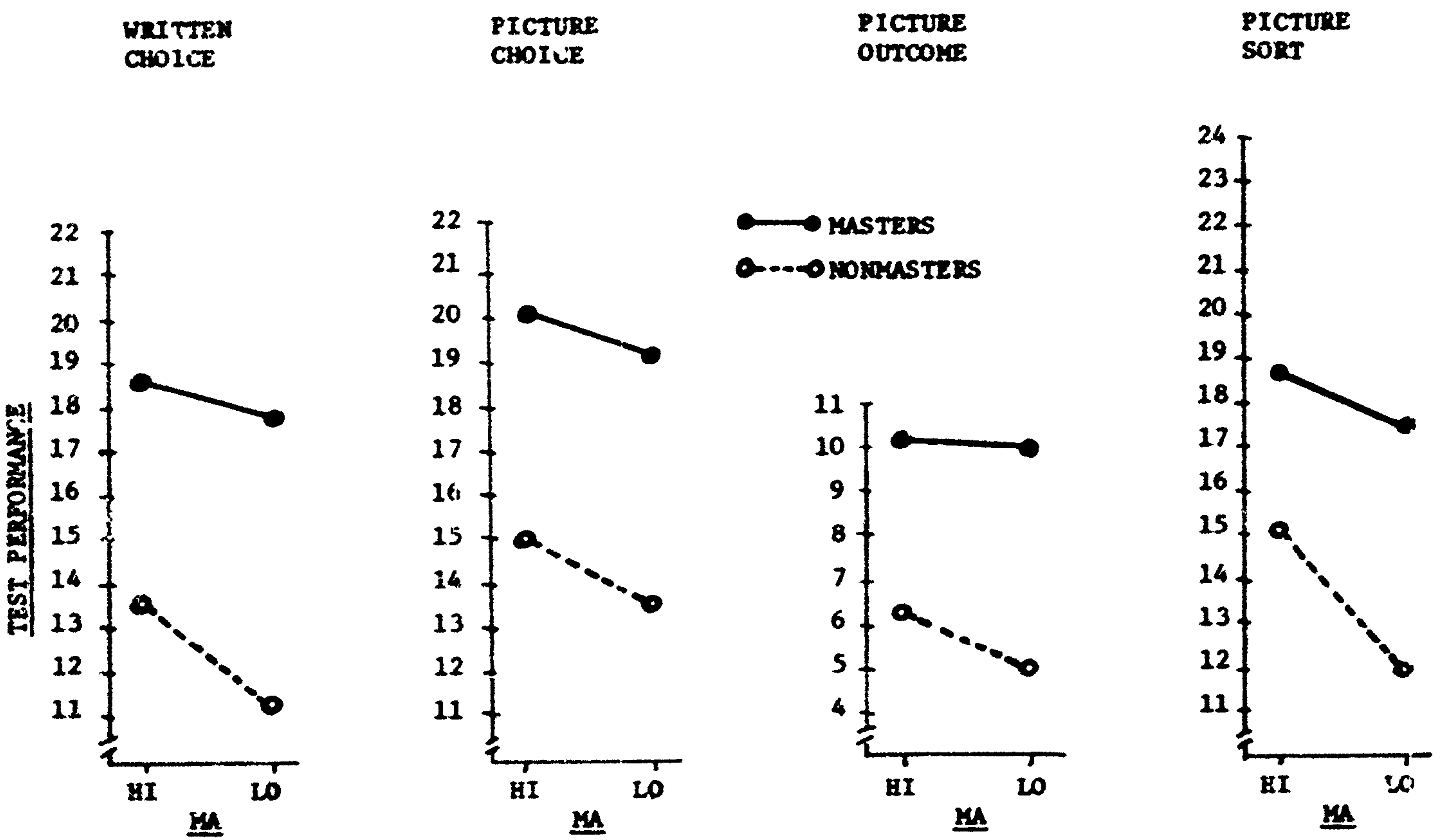


FIGURE 1. Mean performance of masters and nonmasters by mental ability (MA) level for the four knowledge test methods.

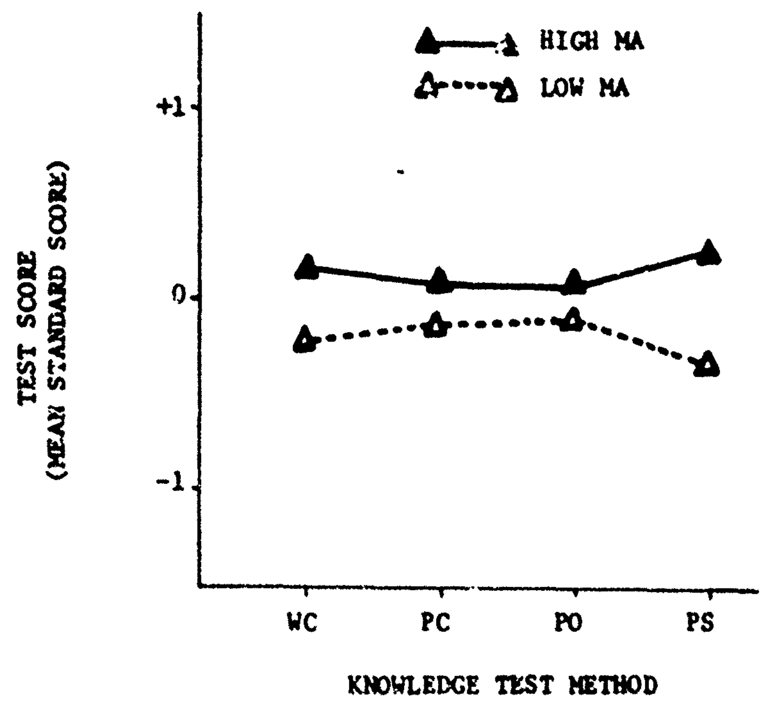


FIGURE 2. Mean standard score performance of high and low mental ability (MA) groups for the four knowledge test methods.

(pass = no more than one item wrong). The results are shown in Table 4 for high and low mental ability groups and for the total sample. With exception of the Picture Outcome method, classification error is somewhat less using the more liberal part mastery criterion on the knowledge tests. Total error tended to run about 25% on the average, reaching a low of 16% for the Picture Choice method with the criterion of part mastery. Of particular interest is the distribution of total error between false-positive and false-negative categories. As the standard for passing a predictor measure is relaxed, the number of false-positives generally increases. The optimal ratio of the two types of error is a moot point, and will depend largely on how test scores are to be used. But if test fairness is the goal, then minimizing the number of false-negatives should be the objective. The relative number of false-negatives, moreover, should be the same for groups differing in mental ability (or any other ability correlated with test score but unrelated to criterion performance). Comparing high and low MA groups we find a small but consistent tendency toward more false-positives among the high MA's, and more false-negatives among the low. This trend was evaluated by Chi-square analysis of the difference in type of classification error between high and low MA groups, and is shown in Table 5 by test method for each standard of test "mastery." Observed Chi-squares were tested at the 10% level of significance, which provides for a conservative decision with respect to accepting the null hypothesis of no difference between groups in distribution of classification error. Type of classification error produced by the knowledge tests does appear to interact with mental ability. Although the number of cases underlying the analysis are too few to warrant firm conclusion, indications are that if one were interested in minimizing the incidence of false-negatives (i.e., the part mastery standard), the Picture Choice method produces the most equitable results for both mental ability groups.

<u>Personal Preferences for Test Methods</u>. *Ss'* opinions of the test methods were solicited after each test was administered and again when all testing was concluded. Responses at the two points in time were similar, so only the final ratings are reported here. *Ss* were asked to rank the five methods (including the hands-on criterion test) from highest to lowest in terms of the question, "Do you think this test is a good way to find out if a soldier can...(e.g., set up a mechanical ambush with a Claymore?") Rankings were done separately for each task. Overall mean preference was highest for the hands-on method of performance testing, as might be expected (Tables 6 and 7). Differences in preference for the four methods of knowledge testing were less pronounced, although the Picture Choice consistently received higher average ranking regardless of the referent task or rating subgroup. Overall, the hands-on method was first, Picture Choice second, Picture Sort third, Picture Outcome fourth, and Written Choice last in average order of preference.

TABLE 4

AVERAGE[a] PERCENT CLASSIFICATION ERROR AS A FUNCTION OF KNOWLEDGE TEST METHOD AND LEVEL OF MENTAL ABILITY

TEST STANDARD	MENTAL ABILITY GROUP	KNOWLEDGE TEST METHOD / CLASSIFICATION ERROR[b] WC FN	WC FP	WC TOT	PC FN	PC FP	PC TOT	PO FN	PO FP	PO TOT	PS FN	PS FP	PS TOT
FULL MASTERY	HIGH	18	05	23	13	07	20	05	15	20	23	03	26
	LOW	27	02	29	25	04	29	16	08	24	33	00	33
	TOTAL	22	04	26	19	05	24	10	12	22	28	02	30
PART MASTERY	HIGH	07	13	20	02	13	15	00	32	32	15	17	32
	LOW	18	02	20	08	10	18	04	20	24	22	02	24
	TOTAL	12	08	20	04	12	16	02	26	28	18	10	28

[a] Averaged over the three tasks.

[b] FN = False Negatives (masters who failed knowledge test)
FP = False Positives (nonmasters who passed knowledge test)
TOT = Total Classification Error

TABLE 5

CHI SQUARE OF THE DIFFERENCE IN TYPE OF CLASSIFICATION ERROR BETWEEN HIGH AND LOW MENTAL ABILITY GROUPS BY TEST STANDARD AND TEST METHOD

TEST STANDARD	KNOWLEDGE TEST METHOD WC	PC	PO	PS
FULL MASTERY	1.33	1.54	4.19*	2.26
PART MASTERY	7.22*	2.49	3.38*	6.30*

*$p < .10$

TABLE 6

MEAN ORDER OF PREFERENCE[a] BY TASK FOR THE HANDS-ON AND KNOWLEDGE TEST METHODS

TASK	HANDS-ON	TEST METHOD WC	PC	PO	PS
TEL	1.14	3.92	3.03	3.58	3.33
AMB	1.25	3.78	2.94	3.72	3.31
RIF	1.08	3.67	3.14	3.47	3.64

[a] The lower the number the higher the preference.

TABLE 7

MEAN ORDER OF PREFERENCE BY SUBGROUP FOR THE HANDS-ON AND KNOWLEDGE TEST METHODS

SUBGROUP	HANDS-ON	TEST METHOD WC	PC	PO	PS
MASTERS	1.00	3.85	3.08	3.38	3.69
NON-MASTERS	1.30	3.83	2.65	3.87	3.35
HIGH MA	1.35	4.10	2.65	3.70	3.20
LOW MA	1.06	3.50	3.00	3.69	3.81
TOTAL	1.19	3.83	2.81	3.69	3.47

Discussion

A number of interesting though tentative findings emerged from this study. The small sample of people and tasks certainly limits generality of the results, and the following interpretation and conclusions should be so tempered.

The data strongly support the hypothesis that performance on manual task procedures is mediated by knowledge. Correlations between task knowledge and task performance were high, particularly for the two procedural tasks with the lowest skill requirements. The correlations reached as high as .75 in spite of the fact that the range of possible test performance seldom exceeded seven points. When performance was aggregated over tasks, the correlations tended to be more on the order of .80.

Substantial differences among methods of knowledge testing were not found. The conventional written multiple-choice test did essentially as well as the pictorially based methods in distinguishing masters from nonmasters. (In this connection, however, it should be noted that test questions were carefully directed at steps necessary in task performance, and did not include those marginally relevant knowledge items often found on such tests.) Failure of the Picture Sort tests to correlate higher with performance was an unexpected result. This method was designed to tap more fully all knowledge aspects of task performance, including recognition of the steps, their correct outcome, and sequence. In so doing, however, it may well have become the most demanding test technique from the standpoint of method-specific mediation requirements; that is, the examinee must first analyze what he does in performing the task, and then synthesize it a step at a time by sorting through a large number of pictures more or less representative of his mental images of the task. That kind of abstract manipulation probably taxes the intellectual and visualization abilities more than we originally anticipated. In support of this speculation, there was some indication that *S*s in the low mental ability group had more trouble with this test method than with others (Figure 1). The written and pictorial multiple-choice tests, though more dependent on literacy, represent a culturally familiar method. The Picture Outcome method appears to be the simplest in the sense of minimizing both literacy and method-specific mediational demands, and is certainly worthy of further study and development as an efficient method of knowledge testing.

Correlations between knowledge and performance were not significantly different for high versus low mental ability *S*s. Yet there was a slight but noticeable trend toward larger correlations within the low mental ability group. The possibility that knowledge measures--including the standard multiple-choice test--are better predictors of task mastery for those of below average mental ability is intriguing. If true, we need to reevaluate the popular notion that knowledge tests of manual performance are unfair to those less apt in the academic skills of reading, writing and symbol manipulation. The notion <u>is</u> probably valid,

but it may be so for reasons quite different than normally offered. Knowledge tests apparently are good predictors of performance on low-skill procedural tasks among people of low to moderate mental ability. The unfairness lies not in the inability of this group to use a knowledge testing medium, but in the tendency of brighter people to over use it. The hypothesis here is that some minimum level of ability, whether innate or acquired, is necessary to handle the symbolic and semantic demands of a knowledge test; but beyond that level, correlated factors such as test-wiseness begin to moderate the true relationship between task knowledge and performance. Two additional features of the data tend to support this speculation: a) higher average knowledge test scores for the high mental ability group, and b) relatively more false-positive errors in predicting mastery among this group.

If one were urged to recommend, on the basis of this study, a method of testing knowledge on low-skill procedural tasks, the Picture Choice would probably have to be named. The data are certainly not conclusive, but this method came the closest to meeting the overall validity criteria: it demonstrated a high correlation with hands-on task performance; the correlation was relatively constant over the range of mental ability; and, the distributions of classification error were more nearly proportional for the two levels of mental ability. Moreover, the Picture Choice method was second only to the hands-on test in examinee preference.

Differential Prediction from an Unexpected Source

Andrew N. Dow, Ed.D.
NETPDC, Ellyson, Pensacola, Florida

During the course of a validity study that is reported elsewhere (Dow, 1977), some previously unsuspected patterns seemed to emerge from the data. If the validation study had been run entirely by computer, and the results picked up from print outs, there is a good chance that these patterns would have remained undetected. However, circumstances were such that it was more convenient to use a small programmable calculator, the Monroe 326, than to arrange programming and runs on the computer. While the Monroe 326 calculates like a computer from a stored program, each bit of data is entered by hand, and each result is read from a visual display and recorded by hand. It was while hand-recording results that the author noticed the emerging patterns.

The test that was the object of the validity study was the communication section of the U. S. Navy's advancement examination administered to candidates for advancement to E-8 and E-9 rates. This section, or rather, these sections, as there is a different one for each paygrade and advancement cycle, are based upon a technique and structure devised by Haney (1953, 1955, and 1958) and called the Uncritical Inference Test.

Haney's Uncritical Inference Tests consist of a short story of 40 to 200 words followed by a series of true-false-? test questions. The directions tell the person taking the test to read the brief story, and accept it as true and accurate. Then, if necessary, re-read the story, and respond to the true-false-? items in order. An answer of "T" means that on the basis of the story, the statement is DEFINITELY TRUE, an answer of "F" means that on the same basis, the statement is DEFINITELY FALSE, an answer of "?" means that, on the basis of the story, you can not be definitely certain about the answer.

During the development of the test, Haney checked both the reliability and uniqueness of the trait that he was presuming to measure. For forms A and B respectively, he found split-half correlations of .762 and .818; when corrected by the Spearman-Brown technique, these increased to .928 and .947. The correlations from the test-retest method were slightly lower, running .67 when form A was followed by form B, and .56 when form B was followed by form A.

Correlations with reading comprehension and general ability tests ranged from .20 to .33. From these figures, it is possible to conclude that the Uncritical Inference Test was measuring a definite, unique, independent trait.

As mentioned in previous works of this author, (Dow 1977, Macaluso and Dow 1969), a communication section was incorporated in each exam for advancement to E-8 or E-9. These communication sections were patterned after the Uncritical Inference Test, and always contained exactly 20 of the True-False-? items. On a purely face-validity basis, a test of this type seems to measure an aspect of being a military supervisor. However, this type of validity is not easily verified; the mathematics is simple, but the criterion is elusive, to say the least. Therefore, the author attempted to validate the communication section against the decisions of the several boards that select the candidates who are to be advanced. It was while he was working on the data for the validation study that the author noted that some ratings tended to have higher scores than others, and that the trends seemed to be rather consistent.

As a first step, a series of correlations were run, using the ratings' mean communication raw score separately for selectees or nonselectees in a paygrade and series as the raw data. For example, one correlation was between the mean scores of series 65, E-8 nonselectees (65-8-NON) and those of series 68, E-8 selectees (68-8-SEL). Another correlation was between series 68 E-8 selectees (68-8-SEL) and series 68 E-9 (68-9-NON) nonselectees. These various correlations are listed in tables 1, 2, and 3.

When tables 1, 2, and 3 are reviewed, it is noted that the coefficients of correlation range from a high of .972 to a low of .527. Table 4 is a frequency distribution of the coefficients found in tables 1, 2, and 3; note that only one coefficient is smaller than .650, and that the median value is .812. The calculated arithmetic mean is .808--rounded to two figures, the mean and median agree at .81. Further note that five of the coefficients are greater than .90; a total of 23 of them are larger than .75.

Before discussing the implications of these rather large correlation coefficients, other facts must be put on record. First, series 65 is the earliest of the three exam cycles, 71, the most recent. Secondly, persons who were not selected (NONs) in a given cycle may participate in the next cycle, and others until he is selected; this means that NONs from a cycle will be included in both the NONs and SELs of the following cycle (series), at the paygrade. The actual percent of overlap of personnel is not known, but is

TABLE 1

X DATA	Y DATA	r_{xy}
65-8-NON	68-8-SEL	.776
65-8-NON	68-8-NON	.972
65-8-NON	71-8-SEL	.819
65-8-NON	71-8-NON	.840
65-8-NON	65-8-SEL	.783
65-8-SEL	68-8-SEL	.681
65-8-SEL	68-8-NON	.768
65-8-SEL	71-8-SEL	.727
65-8-SEL	71-8-NON	.692
68-8-NON	71-8-SEL	.901
68-8-NON	71-8-NON	.905
68-8-NON	68-8-SEL	.763
68-8-SEL	71-8-SEL	.803
68-8-SEL	71-8-NON	.724
71-8-SEL	71-8-NON	.786

Correlation Coefficients Between the Scores Achieved by the Several Specialties in the Various Groups of E-8 Candidates

TABLE 2

X DATA	Y DATA	$\underline{r}_{xy}$
68-9-SEL	71-9-SEL	.844
68-9-SEL	71-9-NON	.812
68-9-SEL	68-9-NON	.756
68-9-NON	71-9-SEL	.898
68-9-NON	71-9-NON	.948
71-9-SEL	71-9-NON	.893

Correlation Coefficients Between the Scores Achieved by the Several Specialties in the Various Groups of E-9 Candidates from Series A and B, Only

TABLE 3

X DATA	Y DATA	r_{xy}
68-8-SEL	68-9-SEL	.527
68-8-SEL	68-9-NON	.771
68-8-SEL	71-9-SEL	.816
68-8-SEL	71-9-NON	.837
68-8-NON	68-9-NON	.878
68-8-NON	68-9-SEL	.721
68-8-NON	71-9-SEL	.862
68-8-NON	71-9-NON	.923

Correlation Coefficients Between the Scores Achieved by the Several Specialties in the Various Groups, Across Paygrades using Candidates in Series A and B, Only

TABLE 4

Correlation Coefficients			N
.950	-	.999	1
.900	-	.949	4
.850	-	.899	4
.800	-	.849	7 ← .812 Median
.750	-	.799	7
.700	-	.749	3
.650	-	.699	2
.600	-	.649	0
.550	-	.599	0
.500	-	.549	1

Frequency Distribution of Correlation
Coefficients Found in Tables 1, 2, and 3

assumed to be rather high--over 50% in some cases. In addition to the overlap caused by reparticipation of nonselectees, some of those selected for advancement to E-8 in cycle 65 caused a further overlap by participating in the cycle 71 E-9 exams. However, because of other data difficulties, there were no correlations involving the 65-8-SEL to 71-9-NON or 71-9-SEL group.

Note that four of the five coefficients greater than .90 are between successive series groups at the same paygrade, with the early group a NON; pairs such as these probably have a high overlap of participants. Even though these are group means rather than individual's scores, these large coefficients probably indicated acceptable test-retest reliability for reasonably equivalent forms of the test. These tests (or sections) are not equated, and their mean raw scores differ noticeably, therefore we cannot call them equivalent. However, because they are of the same structure, and because they do seem to predict one another, they can be called reasonably equivalent forms.

From a different angle, the eight coefficients that are between groups with overlap range from .972 down to .776, with a median of .90. The 21 coefficients that were calculated between non-overlapping groups range from a .923 down to .527 with a median of .78. The differences between the two kinds of groups indicate that the test will predict the score on the retest. These rather large coefficients of correlation show that there is a rating-specific trait that is being measured by the communication subtest of the E-8 and E-9 advancement examinations.

While no detailed data are presented in this paper, the author noted that similar ratings had similar means (high, low, or in between). This opens the possibility that the trait is not exactly rating-predictive, but is related to an occupational area. The possible existence of rating groups or clusters should be investigated.

As the original work was done with a small group of ratings, the study should be repeated for all Navy ratings. Also, the original study used the records only of those candidates that scored high enough to be considered by the selection board, all candidates should be included in this new study.

If further studies do confirm that a reliable measurable trait does exist, and that it is rating specific or interest-area specific, then it becomes necessary to find out whether the trait pre-exists in these persons, or it develops during the years that they have worked in their ratings. Haney (1958) found that experienced policemen scored no higher on the Uncritical Inference Test than police rookies

did. He had assumed that several years of police experience would make a person more critical of the inferences he draws from written material.

If it is found that the trait pre-exists, and is not developed by specific Navy experiences, then it's use as a differential selection device should be pursued. As a rough guess, a predictive instrument would necessarily have more than twenty questions; very likely there should be several short stories, each with 20 to 30 questions.

References

Dow, Andrew N., Twenty Questions as a Predictor of Test Scores and Board Selections for Advancement (Doctoral dissertation, University of Florida, 1977). To be published by University Microfilms.

Haney, W. V. Measurement of the ability to discriminate between inferential and descriptive statements (Doctoral dissertation, Northwestern University, 1953). Dissertation Abstracts, 1954, 14, No. 7037.

Haney, W. V. The Uncritical Inference Test. Wilmette, Ill.: William V. Haney Associates, 1955.

Macaluso, C. J., & Dow, A. N. E-8 and E-9: A new approach. Proceedings of the 11th Annual Conference, Military Testing Association, 1969, 154-163.

Monroe, the Calculator Company. Beta 326 Scientist (Manual of directions for the 326 Scientist programmable calculator.) Orange, N. J.: Monroe, the Calculator Company (Litton Industries.) no date.

ARMY SKILL QUALIFICATION TEST

Dr Clay Brittain

US Army Training Support Center, Fort Eustis, VA

This is the third year running in which we have had a Military Testing Association (MTA) session devoted to SQT. Two years ago, we appeared under the organizational rubric of the Army Enlisted Test Activity (or ETA). Last year, the name was Individual Training Evaluation Group (or ITEG). Now we are Individual Training & Evaluation Directorate (or ITED) of the Army Training Support Center.

One might surmise that name changes as frequent as this, reflects more than the usual bureaucratic proclivity for shifting around boxes and charging labels on an organizational chart and may be indicative of an identity problem, i.e., that we are still trying to find out what we are about. This is not the case. Our goal is essentially the same now that it was last year and the year before that; namely, to implement an approach to occupational proficiency testing which goes beyond the measurement of job-knowledge to the assessment of job-competence. We recognize that such a goal is not unique to ITED, or to the Army. But the SQT program may be unique in the degree of commitment and the magnitude of the effort it represents toward realization of this goal.

A premise which is basic to the SQT program is that testing serves powerfully to stimulate and discipline individual training. Soldiers are strongly motivated to acquire and supervisors to train those competencies which are to be tested. Thus, the SQT program has been designed not simply to serve personnel management needs but to give leverage for focusing and enhancing individual training. The strong emphasis in the SQT program on critical tasks tested realistically derives from this basic aim of insuring that testing is relevant to effective training. The present symposium reflects this point of view.

As a prelude to the symposium, we have had on display for the past two days an exhibit on the SQT program. The display included components of the Individual Training/Evaluation System in which the SQT is embedded. It may be useful here to briefly review this system.

The initiation of SQT development presumes a comprehensive job analysis which identifies the job-tasks critical in an MOS; and for each task a thorough task analysis. The results of job and task analyses are incorporated into a Soldier's Manual. The Soldier's Manual defines for soldiers

the MOS. It lists the tasks which have been identified as critical to the soldier's job and for each task gives a description which includes a statement of the conditions under which the task is to be performed and of the standards which define satisfactory performance. Moreover, the Soldier's Manual delineates the domain of the SQT; i.e., the SQT samples those job tasks listed in the Soldier's Manuals. Thus, the Soldier's Manual, in effect, says to the soldier "If you are to be competent in your MOS, these are the tasks you must be able to perform. In assessing your occupational competence, we will not go outside the tasks listed here". The Soldier's Manual must be in the hands of soldiers at least six months prior to the time they take the SQT.

Sixty days or more prior to the SQT, the soldier receives an SQT Notice. This document identifies the specific tasks to be tested and for each task specifies the SQT component. That is, the soldier is told whether the task is to be tested in the written component (WC), the hand-on component (HOC), or the performance certification component (PCC). In selecting tasks for the SQT, the aim is to test those tasks on which performance deficiencies are most prevalent and/or most serious. The purpose in issuing an SQT Notice in fairly close proximity to the test period is to focus individual training efforts upon those tasks most in need of training.

SQT results are reported to soldiers and to various echelons of command. The soldier gets an individual soldier's report (ISR) which identifies the tasks on which he (or she) failed and gives an overall SQT score. Commanders - from the battalion to the major command level receive reports which show in aggregate how soldiers in their command did on each task tested; i.e., pass/fail percentages for each task. The aim is to provide to each level of command information useful in managing, supporting, and facilitating individual training.

The SQT program is being implemented on a schedule which will be completed in about two years. Skill Qualification Testing for record began last April with the testing of soldiers in Career Management Field (CMF) 11 (Maneuver Combat Arms). The testing of soldiers in CMF 16 (Air Defense) and CMF 95 (Military Police) began in July. The testing of soldiers in CMF 76 (Supply) began this month. With the phase-in of additional CMF each quarter, SQT will have been implemented for all enlisted CMF in the Army in the first quarter of FY 1980.

With this brief background statement on the SQT program, we now turn to the present symposium, which falls somewhat logically into two sections. In the first two presentations, SQT developed by the Military Police School and by the Air Defense Artillery School will be described and discussed. Following these presentation with a slight shift in perspective, we will present and examine some early SQT results, discuss our experiences with performance testing, describe the training of SQT developers, and reflect on some of our problems and lessons learned.

Use of Video-Tape for the Military Police Skill Qualification Test (SQT)

Daniel E. Spector, PhD, Chief, SQT Development Branch
EPMS Division, Directorate of Training Developments,
USAMPS/TC, Fort McClellan, Alabama 36205

Many military police tasks involve quick decisions in reaction to what a policeman or woman sees or hears. Conventional paper and pencil tests have measured ability to perform these tasks by presenting a detailed written situation and requiring the examinee to choose the appropriate action for responding to the situation. As the task cues are visual and/or audial, presentation of a "word-picture" involves a serious drop in test fidelity. Responding to a written situation is simply not the same as responding to a dynamic visual problem presented with sound. Moreover, the test developer can never be sure the word picture conjures up the same image in every examinee's mind. Finally, a written test cannot require a real time response. The examinee can ponder the situation or go on to another question in hope of a hidden cue or flash of brilliance. There is no such luxury in most real world police tasks.

The US Army Military Police School is overcoming some of these problems through use of video-tape in the Skill Qualification Test (SQT) program. If a task involves response to visual or audial cues, these can be presented to examinees on television through use of video-tapes. Task fidelity is thereby greatly enhanced. All examinees see and hear a situation very much as if they were on the job. As they all see the same thing, the test developer no longer has to worry about a "word-picture" meaning different things to different examinees. Perhaps the greatest benefit of this testing mode is the requirement for a real time decision. To paraphrase Omar Khayyam, the moving picture moves on and only the test monitor can move it back. The examinee must make a decision quickly; in but a few seconds another problem will be presented for a decision, very much like in real life.

The 1977 SQT for military police uses video-tape to test five tasks. Perhaps the most obvious candidate for a video-tape test is the task of enforcing traffic regulations. Task cues are entirely visual and dynamic. Written description of possible traffic violations, even when augmented by pictures or illustrations, cannot capture the task very well. The video-tape can do much better. The examinee is told to imagine himself, or herself, behind the wheel of a patrol car. The camera is to be the eyes of the examinee. The test then presents ten traffic scenes. For each, the examinee must decide whether a violation has occurred, and, if so, just what it is. We think this test comes very close to real world performance. It certainly has greater task fidelity than a paper and pencil test, while avoiding the obvious problems of administering a fully hands-on test.

Another task in our video-tape test is use of observation and description techniques. The examinee is given descriptions of several people and vehicles. Again, the camera serves as the eyes of the examinee. The test then shows people milling around a building, going in, or coming out. The examinee has to select those that match the descriptions. For vehicle identification the examinee is "driven" through a parking lot and forced to pick out those cars that match the descriptions. Although this task could be tested in a paper and pencil mode if adequate pictures were given, there would be some loss of task fidelity. The video-tape presents a dynamic situation and forces a real time decision.

We are also using video-tape to test the tasks of warning suspects of their rights and receiving and processing offenders. The rights warning is not as simple as it seems. The military police must know when the warning is necessary; this is especially difficult in interviewing a witness who says something that may make the police think they have a suspect. There are also problems involving civilian as opposed to military suspects and with determining when legal counsel is necessary. Our video-tape presents five vignettes. After each, the examinee must determine what, if anything, was done wrong. The same thing is done with the receive and process offenders task. In these tests the video-tape presents the audial cues of the task as well as the visual ones. A paper and pencil test could not do this.

The other task in our video-tape test involves recording data in the MP notebook. This is extremely critical as the notebook is the basis of subsequent reports, and adequate notes could be critical to the outcome of a court case. In this task both visual and audial cues are important. The examinee is told that he and his partner are investigating a crime. The examinee is to record the results of an interview in the notebook and then walk around the crime scene, noting possible clues and evidence. The camera is the eyes of the military police.

This task presented two technical problems. In real life the military police can ask a question over if the answer is not clear. The test cannot allow this, but we do have the military police repeat the information as it is being noted. We will be looking at this carefully to see if this serves to overcue the examinee. Another problem was when to ask questions about the task. We wanted to insure the examinee used the notes, not just his or her memory. To force this the note taking part of the test appears at the very beginning of the video-tape test. The examinee then takes the other parts of the test. At the end of this, about 45 minutes later, the examinee must answer questions about the notes. We think that the intervening tasks will erase or confuse the examinee's memory, thus forcing reliance on the notes.

(Excerpts of the video tape test will be shown here).

The military police video-tape SQT was an "in-house" Army project. It involved no contract. No purchase of equipment was required. There was minimal TDY expense, as the Army has complete video-taping facilities at Redstone Arsenal, only 100 miles from the MP School at Fort McClellan. The tests were written by noncommissioned officers, subject matter experts in the tasks. They were assisted by test development and video-tape production personnel at the two posts involved. All acting was done by Army personnel.

In short, this sort of test may be more difficult to develop than a paper and pencil test, but it is certainly not beyond the capabilities of the average service school. The MP School plans to make heavier use of video-tape in SQT for 1978. We will use this mode for testing our criminal investigators and corrections specialists; we will also expand its use for the military police.

This testing mode may also prove useful for testing resident students. It can present the student with a more realistic problem than a paper and pencil test. Furthermore, it's a lot more fun to develop. What more can one ask?

SQT IMPLEMENTATION: EARLY RESULTS, EARLY LESSONS

Clay V. Brittain
US Army Training Support Center

INTRODUCTION

In this presentation, I will describe some properties of SQT so as to convey a sense of the rationale of skill qualification testing, and within the content of this rationale analyse early SQT results. I will also allude to some of our experiences in implementing the SQT program.

A basic concept of the SQT approach to occupational proficiency testing is job-task. The primary aim is to test the ability of soldiers to perform job-tasks which have been identified as critical to their MOS. At the risk of shunting aside a plethora of difficult questions, I am going to assume that there are no serious problems in identifying and meaninfully defining job-tasks or in determining task criticality.

The basic module of the SQT is the scorable unit (SU). A task usually is represented on the SQT by one SU, although complex tasks might be represented by two or three SU. The SU confronts the soldier with a task, or a series of subtasks which more or less faithfully capture the essential features of the job task which it represents. The SU might be regarded as a sub-test, and the SQT as an aggregate of these subtests. The SU provides the basis for categorizing soldiers as "performers" or "non-performers" of the referent task. Soldiers are scored "GO" or "NO-GO" and receive a score of "1" or "0" on each scorable unit. The total SQT score is derived from an aggregation of these SU unit scores; i.e., the SQT score reflects the proportion of SU in which the soldiers scored "GO". Thus, the SQT scores run from a minimum of 0 to a maximum of 100.

For the individual soldier, the SQT score yields one of three possible outcome: (1) failure, (2) verification (or lower passing scores) and (3) qualification (or higher passing scores). The higher-passing score is 80 or higher and is taken to indicate that the soldier is technically qualified for award of the skill-level next above the one presently held. The lower-passing score is 60-79, and is taken as indicative that the soldier is technically competent in his present skill-level. The failing score is 59 or less and is taken to indicate that the soldier lacks the required technical competence at his present skill-level.

These SQT scores are highly significant for the soldier's career On the one hand, a qualification score on the SQT is prerequisite to award of the next higher skill level and to the soldier's eligibility to compete for promotion. On the other hand, after two successive SQT failures, the soldier is vulnerable to adverse personnel actions.

Responsibility for implementing the SQT program has been assigned to the Army Training and Doctrine Command (TRADOC). This is not fortuitous. The SQT is training-oriented. The quest for training-relevance dictates much of the logic of the system. In selecting tasks for testing, the aim is to select those tasks which need to be made the focus of training. In reporting SQT results, the aim is to provide information for managing and directing training. But at the same time, SQT scores provide a highly important input for the personnel management system.

The import here is that the SQT must serve two masters: Training management and personnel management. This can impose divergent informational requirements on the SQT. Resolving the competing demands of the trainers and personnel managers has been one of the problem areas not easy of satisfactory resolution. But if this complicates life for us, it seems a necessary complication. If the SQT is to provide the leverage for focusing and directing individual training efforts, then the SQT score must have career-relevance. It must make a difference in terms of the soldier's movement up the career-ladder.

SHAKEDOWN TESTING

Prior to testing for record, which began last April, SQT were administered on a trial basis to soldiers in four MOS. The purpose was to shakedown the SQT system. The scores were not entered into the records of the soldiers tested. The results of the shakedown testing will not be reported in detail here. But, to summarize briefly, of the soldiers tested, 76 percent failed, 23 percent verified their present skill level, and less than one percent qualified. There many be many reasons for these predominantly low scores, but I assume that to some extent they reflect the fact that the testing was not for record. Promotion, classification, or retention in the Army were not at stake for the soldier.

Of the lessons learned from the shakedown testing, let me cite three. 1) Hands-on performance testing. Although data on soldier reactions to the SQT were not systematically collected, there was quite a lot of informal feedback which indicated that the soldiers reacted very positively to the hands-on component. But at the same time, we became newly aware of the just how demanding it can be to develop even a relatively simple performance test. We spent long hours with the SQT developers in the test development agencies in specifying performance measures in the detail and clarity required for uniform and objective scoring. Also the formulation of instructions on setting up the test station and administering and scoring the test was difficult and time consuming. But, nonetheless, one of our major aims is to expand the HOC. 2) Logistics. Another problem about which the shakedown made us wiser, or at least more sensitive, had to do with the logistics of supporting hands-on performance

testing. For example, if soldiers are to be tested on the ability to toss hand-grenades then it is necessary that "practice" hand-grenade be available not only for testing, but for training. In our guidance to the schools, where the SQT are developed, we have become emphatic on this point, i.e., be sure that provision is made to insure the availability of supplies and equipment required in the hands-on component. Be very cautious in your assumptions on availability of equipment. 3) A third lesson had to do with errors in filling out SQT answer sheets. The SQT has three components. There is a separate mark-sense answer sheet for each component. Whatever else might be said about this set-up, it seems to be a sensitive and powerful test of the soldiers' clerical ability. Almost 100 percent of the SQT answer sheets returned to ITED for scoring contained at least one clerical error. It is a profound tribute to the Field Services Staff and the Data Processing folks at ITED that virtually all these answers sheets were finally scored. We have succeeded in significantly reducing this error rate. But forms simplification remains one of our major aims.

TESTING FOR RECORD

Testing for record began last April with the testing of soldiers in CMF 11 (Maneuver Combat Arms). SQT 2, 3, and 4 were administered to soldiers in the following MOS: 11B (Infantryman), 11C (Indirect Fire Infantryman), 11D (Armor Reconnaissance Specialist), 11E (Armor Crewman).

Before presenting and discussing results from this first round of testing, it is necessary to turn aside briefly to clarify two points: (1) the scheme for numbering SQT, and (2) the use of tracks in SQT. First let me comment about the numbering of SQT.

Numbering of SQT. The number of an SQT reflects the skill-level of soldiers who take it. There are five skill-levels in the Army enlisted MOS structure and these are articulated with pay grade as shown in slide 1.

(Slide One)

Also shown here is the relationship between skill-level and SQT number. SQT 2 matches skill-level 1, SQT 3 matches skill-level 2, SQT 4 matches skill-level 3, and SQT 5 matches skill-level 4 and 5.

The logic in having the SQT one number higher than the skill-level is that the soldier takes the SQT as a means of qualifying for the next higher skill level. For example, skill-level 1 soldiers take SQT 2 as a means of qualifying for skill-level 2 and skill-level 2 soldiers take SQT 3 as a means of qualifying for skill-level 3. The logic of this system of numbering SQT is the same as the logic for calling them skill-qualification tests.

SQT Tracks. Now with the SQT numbering scheme hopefully clarified, let me comment about SQT tracks. It is not unusual that soldiers holding the same MOS and assigned to duty positions in that MOS actually work with different types of equipment or perform different types of duties. For example, soldiers in a given MOS may work either with equipment A or equipment B, but virtually never with both A and B.

This poses a problem in proficiency testing. Should soldiers who work only with equipment A be held accountable for proficiency in equipment B, and vice-versa? Consistent with the aim of testing only those competencies which are directly relevant to the soldier's job, the resolution of this question in the SQT program has been a policy decision to allow MOS proponent agencies to track SQT: i.e., to have the SQT made up of a core of tasks on which all soldiers in the MOS are tested plus two or more parallel segments, each made up of tasks pertinent only to certain equipment systems and/or duty positions. The proponent agency decides on the number of tracks and specifies the rules for assigning soldiers to tracks.

Infantry Results. Now, after this digression to talk about SQT numbering and the policy on tracking SQT, let me return to the main thrust of the presentation and discuss some results of the first round of skill qualification testing. The presentation here is limited to MOS 11B (Infantryman) and MOS 11C (Indirect Fire Infantryman). SQT results for soldiers in these two MOS are summarized in Slide two.

(Slide Two)

The slide shows - by skill level - the percentages of soldiers who (1) qualified for the next higher skill level (score "GO" on 80 percent or more of the tasks tested), (2) verified the present skill level (scored "GO" on 60 percent to 80 percent of the tasks), and (3) failed to verify the present skill level (scored "GO" on less than 60 percent of the tasks tested).

As shown here, the 11C soldiers scored lower than 11B soldiers, with the 11B vs 11C differences being biggest at skill level 1 (SQT 2) and smallest at skill level 3 (SQT 4). My major purpose now is to examine the meaning of these differences, limiting the analysis to SQT 2 of MOS 11B and 11C.

It is pertinent here, to note that the 11C test had two tracks: Track 1 for soldiers who work with the 81 MM mortar and Track 2 for soldiers who work with the 107 MM (4.2 inch) mortar. As a group, soldiers taking Track 1 consistently scored slightly higher than soldiers taking Track 2. This superior performance of soldiers working with the 81 MM mortar was reflected in higher pass rates on tasks in the non-tracked portions of the SQT as well as the tracked portions. In comparing 11B and 11 C SQT performances, the 11C scores will be from soldiers taking Track 1 (i.e., soldiers working with the 81 MM mortar). The differences within MOS 11C between soldiers working with the 81 MM mortar and soldiers working with the 107 MM mortar is a separate issue.

It is consonant with the logic of skill qualification testing to say that, in the absence of plausible competing interpretations, low SQT scores are indicative of the need for training. The implication here then is that MOS 11C soldiers are not as well trained as MOS 11B soldiers in the duties of their respective MOS. But there are competing interpretations which need to be examined; namely, interpretations in terms of test-practice effects and of difficulty.

Test Practice Effects. The role of test practice as a credible interpretation is based upon the fact that SQT 2 and 3 of MOS 11B were administered Army-wide last year as part of the shakedown testing. Thus, many of the 11B soldiers were taking the SQT for the second time. This was not true of 11C soldiers.

In connection with the question of test practice effects, it is of interest to examine SQT shakedown scores in relation to the "For record" SQT scores. The scores are summarized in Slide Three.

(Slide Three)

As shown here, the 11B soldiers did considerably better the second time around. No soldier earned qualifying scores in the shakedown as compared with 22 percent of the soldiers qualifying in the record testing. In the shakedown testing, the failure rate was 82 percent versus a failure of 31 percent in the testing for record.

Although consistent with a "practice-effects" interpretation, it is not clear that these differences really reflect test practice effects. Since the "for record" SQT scores had career-relevance for the soldier which the shakedown scores lacked, it is plausible to attribute the higher scores in the "for record" testing to stronger motivation to do well on the test. It may be more relevant here to compare 11B and 11C scores on those tasks which are common to the two SQT. First, let us examine the Performance Certification Component (PCC) and the Hands-on Component (HOC).

On both the 11B and 11C SQT, the PCC included an arms qualification test (Qualify wit M16A1 Rifle) and the Advance Physical Fitness Test (APFT).

(Slide Four)

The rifle qualification test is summarized in Slide Four. The soldier was scored as follows:

Failure to Qualify	0 Scorable Unit
Marksman	1 Scorable Unit
Sharpshooter	2 Scorable Units
Expert	3 Scorable Units

(Slide Five)

The APFT is summarized in Slide Five. It was scored as follows:

Failure to Qualify	0 Scorable Unit
300-399 Points	1 Scorable Unit
400-449 Points	2 Scorable Units
450-500 Points	3 Scorable Units

(Slide Six)

Scores on the arms qualification test are summarized in Slide Six.

(Slide Seven)

Scores on the Advanced Physical Fitness test are summarized in Slide Seven. MOS 11B and 11C soldiers performed comparably on these tasks.

(Slide Eight)

In addition to the rifle qualification and physical fitness tests, the 11C PCC included a Gunner's Exam on which soldiers were scored as unqualified with no scorable units, or as second class, first class, or expert gunner and credited with 1, 2, or 3 scorable units. The scores are summarized in Slide Eight. The percentages of soldiers qualifying at these three levels was lower here, but mainly reflected a higher percentage not rated (i.e., more soldiers who had not taken the Gunner's exam).

(Slide Nine)

The hands-on component included six tasks, five of which were common to the two SQT. The HOC results are summarized in Slide Nine. The two groups performed very comparably on the five common tasks. On the sixth task, which was unique to each SQT, 69 percent of the 11B soldiers scored "GO" and 62 percent of the 11C soldiers.

The results of performance on the PCC and HOC do not give evidence of practice effects. On those PCC and HOC tasks common to the two SQT, the 11B soldiers did no better than 11C soldiers. But it is of interest here, to point out that the SQT Notice which soldiers get at least 60 days prior to SQT, gives very detailed and complete information on the PCC and HOC. Soldiers could get a comparable level of information about the WC only by actually seeing the test booklet. Thus, it would seem reasonable to expect that whatever advantage might have accrued to 11B soldiers for having taken the SQT before would not be as strongly reflected in performance on the PCC or HOC as on the WC.

(Slide Ten)

Of 35 scorable units in the WC of these SQT, 15 were common to the two SQT. The 11B and 11C pass rates on these 15 tasks are shown in Slide Ten.

MOS 11B pass rates were <u>higher</u> on eight and <u>lower</u> on six of these tasks. Pass rates were equal on one task. The differences were generally modest with the most notable exception occuring on Task 071-11A-1501 (call for/Adjust Indirect Fire, using Grid Coordinate Method of Target Location and Bracketing Method of Adjustment). Thirty-five percent of 11B soldiers scored "GO" on this task as compared to 53 percent of 11C soldiers.

MOS 11B soldiers did not perform better tha 11C soldier on the common written SU. Thus, the data here do not support the argument that 11B SQT scores were higher than 11C because of test practice effects. What the data do indicate is that 11C soldiers most frequently failed those tasks unique to the 11C SQT.

DIFFICULTY. The three psychologists at ITED who are most familiar with SQT 11B2 and SQT 11C2 agreed in the judgement that the 11C test confronts the soldier with questions which tend to be cognitively more difficult than 11B questions. In order to put these impressions to a more critical test, the two SQT were systematically examined with respect to readability and content.

In assessing readability, it is recognized that conventional readability formulas such as the fog-count probably are not fully applicable to the SQT. These formulas utilize sentence length and word length as indicators of readability. In a document which uses technical terms familiar to the audience, word length may cease to be effective as a readability index. For example, the use of "camouflage" in an SQT probably does not have the significance that a word of this length would have in nontechnical writing. However, it seems reasonable to assume that readability analysis is no less applicable to one of these SQT than the other. Thus, differential readability scores derived from the tests presumably would be indicative of actual differences in readability. Two different readability formulas were applied. The fog-count estimated the readability of both SQT to be at about the seventh grade level. A readability formula proposed by Fry ielded slightly different estimates; about the sixth grade level for SQT 11B2 and one grade higher for SQT 11C2. The difference here reflects sentence length. The situational descriptions and item stems included slightly longer sentences in SQT 11C2 than in SQT 11B2.

Along the same line, it is of interest that the reading load as reflected in the total number of words in the situational descriptions and item stems is greater by about 15 percent in SQT 11C2 than in SQT 11B2. The 11C2 written scorable units on the average are slightly longer in terms of number of items. Of the 20 scorable units which are unique to the 11B SQT, one-half have four or less items. Of the 20 SU unique to the 11C SQT, one-half have six or more items.

In summary, then, there are indications that SQT 11C2 may impose slightly higher demands on the soldier in terms of reading burden.

In relation to the question of test difficulty, it is relevant also to ask whether the 11B and 11C SQT are different in terms of the kinds of tasks required of the soldier? The distinction between "written

performance" and "performance based" testing is relevant here. ITED guidance on SQT development characterizes written performance (WP) testing as that which requires the soldier to perform a task (or task segment) essentially as it would be performed on the job; whereas performance based (PB) testing requires the soldier to answer questions about task performance. Two ITED psychologists analyzed SQT 11B2 and SQT 11C2 in terms of WP and PB testing, and categorized test items as WP or PB and scorable units as WP, PB, or mixed. This analysis revealed no differences between SQT 11B2 and SQT 11C2. Of the 20 written SU unique to the 11B SQT, 10 were identified as WP or mixed. Of the 20 written SU unique to the 11C SQT, 11 were identified as WP or mixed. Thus, the two SQT were very comparable in regard to utilization of WP and PB testing.

In another analysis, we grouped SU into the following categories based upon the types of behavior required of the soldier.

- Word recognition
- Picture recognition
- Chart reading
- Mathematical computation

It was assumed that the SU in the first two categories are generally easier than SU in the last two categories, and pass rates are consistent this assumption, i.e., lower on SU involving chart reading and mathematical computation than on SU requiring word or picture recognition. Of the 20 written SU unique to the 11B SQT, only one involvled mathemaitcal computation or chart reading. Of the 20 written SU unique to the 11C SQT, nine involved mathematical computation and/or chart reading.

This analysis suggests that the 11C SQT may have been a more difficult test than the 11B SQT. It is pertinent to ask whether this is spurious or is reflective of "real world" differences between these two MOS. Are tasks which 11 soldiers are required to perform cognitively more complex than tasks 11B soldiers are required to perform? This question cannot be answered from the present data, but it is of interest that the more difficult tasks on the SQT were those unique to MOS 11C. These tasks had to do mainly with gunnery operations.

Now let me briefly summarize. SQT scores for two MOS were compared and found to be quite disparate. As compared to their counterparts in MOS 11B, soldiers in MOS 11C scored substantially lower. It is consistent with the logic of skill qualification testing to interpret these results as indicating differential need for training. Competing interpretations have been examined. The analysis argues against attributing the differences to test practice effects. There are indications that the 11C SQT may have been more difficult, but not necessarily in a spurious way. The 11C soldiers had their highest failure rates on tasks specific to the MOS. These tasks had to do mainly with mortar operations.

In concluding, a major aim in the SQT program is the identification of performance deficiencies which need to be made the focus of training. The present analysis provides an early indication of the realization of this aim.

PAY GRADE, SKILL LEVEL, AND SQT NUMBER

Pay Grade	Skill Level	SQT
E-1 - E-4	1	2
E-5	2	3
E-6	3	4
E-7	4	5
E-8 - E-9	5	5(?)

Slide 1

SQT SCORES FOR MOS 11B AND 11C: PERCENTAGE OF SOLDIERS QUALIFYING, VERIFYING AND FAILING

	SQT 2		SQT 3		SQT 4	
	11B	11C	11B	11C	11B	11C
QUALIFYING	22%	4%	27%	10%	26%	23%
VERIFYING	46	26	49	39	49	44
FAILING	32	70	24	51	25	33

SLIDE 2

SQT 2 SCORES, MOS 11B: SHAKEDOWN VERSUS FOR RECORD T

SHAKEDOWN

	SHAKEDOWN	RECORD
QUALIFYING	0%	22%
VERIFYING	18%	46%
FAILING	82%	32%

SLIDE 3

Task Number	Task Title	Scorable Units
(1) 071-11A-2005	Qualify with an M16A1 Rifle (Engage Targets with an M16A1 Rifle) (Specific Course of Fire to be Determined by Local Conditions) (SL 1)	A. Fail to Qualify: 0 Units B. Marksman: 1 Unit C. Sharpshooter: 2 Units D. Expert: 3 Units

SLIDE 4

(2) 071-11A-0201	TAKE THE APFT (MAINTAIN PHYSICAL FITNESS) (SL 1)	A. FAIL TO QUALIFY: 0 UNITS B. 300-399: 1 UNIT C. 400-449 2 UNITS D. 450-500 3 UNITS

SLIDE 5

TASK: QUALIFY WITH M16A RIFLE

	11B	11C
NOT RATED	18%	17%
FAIL (0 UNITS)	1	1
MARKSMAN (1 UNIT)	30	32
SHARPSHOOTER (2 UNITS	69	71
EXPERT (3 UNITS)	45	46

SLIDE 6

TASK: APFT (MAINTAIN PHYSICAL FITNESS)

	11B	11C
NOT RATED	13%	9%
FAILED (OUTFITS)	2	2
300-397 (1 UNIT)	85	89
400-447 (2 UNITS)	66	68
420-200 (3 UNITS)	24	30

SLIDE 7

TASK: GUNNER'S EXAM

NOT RATED	39%
FAILED (0 UNITS)	3
2D CLASS GUNNER (1 UNIT)	22
1ST CLASS GUNNER (2 UNITS)	46
EXPERT GUNNER (3 UNITS)	30

SLIDE 8

HANDS-ON COMPONENT, SQT 2: MOS 11B VS 11C

TASKS	PERCENT GO 11B	11C
1. INSTALL/RECOVER AN ELECTRICALLY ARMED CLAYMORE MINE; FIRE CLAYMORE (071-11A-4501 & 4502)	75%	73%
2. PREPARE AN M72A2 LAW FOR FIRING; ENGAGE TARGETS (071-11A-2201)	89	87
3. PREPARE/OPERATE TEACTICAL FM RADIOS (071-11A-0930)	85	85
4. ENGAGE THE ENEMY WITH HAND GRENADES (071-11A-4402)	87	86
5. PUT ON A PROTECTIVE MASK (071-11A-0103)	87	86
6. LOAD, REDUCE A STOPPAGE, UNLOAD, AND CLEAR A M60 MACHINE GUN (071-AAB-3001)	69	
7. PREPARE M16 PLOTTING BOARD FOR OPERATION AND DETERMINE INITIAL FIRING DATA FOR MORTARS (071-11C-1601 &1602)		62

SLIDE 9

WC Scorable Units shared by SQT 11B2 & SQT 11C2.

TASK	PERCENT GO 11B	PERCENT GO 11C
1. Apply the four life-saving measures (071-11A-0001)	53%	51%
2. Maintain Protective Mask and Accessories (071-11A-0101)	93	92
3. Take Cover as Protection against NBC Hazards (071-11A-0104)	88	84
4. Administer Antidote to a Nerve Agent Casualty (901-11A-0150)	53	47
5. Move under Direct Fire (071-11A-0502)	77	79
6. React to Flares (071-11A-0511)	61	58
7. Use Challenge and Password (071-11A-0801)	53	51
8. Identify Threat Vehicles and Weapons (071-11A-0806)	86	87
9. Install/Operate Field Telephones (TA-1&TA-512) (071-11A-0902)	34	35
10. Call for/Adjust Indirect Fire (071-11A-1501)	35	53
11. Load, Reduce a Stoppage, Unload & Clear an M16A1 Rifle (071-11A-2003)	54	54
12. Zero an M16A1 Rifle (071-11A-2004)	51	50
13. Engage Targets With an M203 Grenade Launcher & Apply Immediate Action to Reduce a Stoppage (071-11A=2104)	56	48
14. Maintain a Caliber .45 Pistol & Ammunition (071-11A-2401)	46	53
15. Perform an ESC Inspection on a Wheeled Vehicle (071-11A-6005)	82	83

SLIDE 10

PERFORMANCE TESTING IN THE SQT

Dr Frank M. Aversano
Mr David H. Poole

INTRODUCTION

As indicated already, the SQT may consist of three parts: A written component (WC), a hands-on component (HOC) and a performance certification component (PCC). The written component is a paper and pencil test. The hands-on test, as its name suggest, tests the soldier by observing his/her performance on job equipment or simulators. The performance certification component is a performance test given on the job. The PCC is used when testing a task in the HOC is impractical or requires too much time, equipment or other resources. The focus of this paper is the hands-on component.

In the SQT program we have proceeded somewhat cautiously in regard to performance testing, and so far the hands-on component has been relatively small. But this is changing. Eventually, the SQT in some MOS may be made up mostly of hands-on testing.

But the first steps into Army-wide performance testing have been taken. This is an initial and somewhat limited progress report on our early experiences. In no way should the paper be considered as the last word on performance testing in the SQT program, or even the first word in any official sense.

In organizing our comments here, we draw upon Fitzpatrick and Morrison's chapter on performance and product evaluation in Educational Measurement (1971) edited by Thordike. These writers discuss performance testing in terms of stimulus aspects and response aspects. The HOC will be discussed here in terms of test instructions, stimuli presented to the examinee, stimuli surrounding the examinee, response aspects, and test administration. We will also present an example of a hands-on score sheet and touch upon the topics of test reliability and validity.

INSTRUCTIONS

All hands-on tests have extensive instructions which must be written for four persons who are involved in the administration of the hands-on test. These persons are the Test Control Officer (TCO), Test Site Managers (TSM), scorer and examinee. The Test Control Officer is the individual responsible for the administration of SQT in the field. He/she is the point of contact between the field and ITED. It is important that instructions be written in sufficient detail to allow the TCO to organize for the administration of the HOC. This places a burden on test developers to consider all aspects of the testing situation.

The Test Site Manager (TSM) is appointed by the TCO to supervise the administration of the HOC. The TSM is responsible for the equipment used in the test, the set-up of the test site, the rotation of soldiers through the test site and the training of scorers.

Scorers are those individuals who actually administer the HOC and score soldiers' performance.

The importance of explicit instructions, training and rehearsal for the scorers cannot be overemphasized. In some cases TV tapes have been developed to show the scorers how to score the HOC. Examples of both correct and incorrect performance are demonstrated with concurrent marking of the score-sheet. In addition, the scorers are required to practice scoring using the scoresheet. In a simulation of the test, scorers play the role of examinees and intentionally perform incorrectly while fellow scorers evaluate their performance. All scorers must play the role of examinee and scorer before being permitted to score on the actual test.

The fourth and perhaps most important role in the hands-on test is that of examinee. The problems that deal with examinee instructions tend to be of two types. In the first type the instructions are so vague that the examinee does not know what behavior is required to demonstrate mastery of the task. In the second type, instructions are too detailed. Instructions are too detailed when they cue the soldier to the desired behavior. For example, there is one task that requires the soldier to fire the LAW (Light Anti-Tank Weapon). One of the performance steps is to inspect the LAW before firing to insure that the weapon is free of cracks. If the scorer tells the examinee to inspect for cracks, he/she has cued the soldier to correct behavior. Excessively detailed and vague instructions can be reduced by outside review and trail runs.

Instructions are also clarified for the examinee by releasing the scoresheet 60 to 90 days prior to the test in the SQT notice. This procedure has a number of advantages. In addition to providing the examinee with information about the task to be tested, it may motivate the examinee to prepare for the test and directs the soldier's training to critical tasks. A disadvantage of this procedure is that the soldier may have a tendency to practice only those steps that are on the scoresheet. Sometimes the scoresheet may not be inclusive because of the difficulty encountered in reliably observing some performance step. In some cases this has been solved by putting the performance step that is difficult to observe on the scoresheet but not scoring the performance of the step during the test. In this way, training is maintained as well as reliable measurement.

Stimuli Presented to the Examinee.

The stimuli presented to the examinee during the test should approximate those stimuli found on the job. The stimuli on the hands-on test can come in three forms: media, simulators and actual equipment. Each one of these forms is appropriate in different situations. Media, for example, are appropriate when no direct intervention is required on the part of the examinee and when the job stimuli are transitory. For example, a Military Policeman might only see a suspect or stolen vehicle for a few seconds before a decision to act must be made. Similarly, a soldier may only catch a brief glimpse of a tank or aircraft before he/she must decide if it is part of some threat force. However, when intervention is required by the examinee, media usually are not appropriate.

Simulators also present a useful alternative for the test developer. When the use of actual equipment is impractical or impossible because of expense, expendability of equipment, safety reasons, and the large amount of soldiers being tested, simulators provide a reasonable alternative for the test developer. One problem encountered with simulators concerns the degree of realism in the simulator. Many developers believe that simulators are an inadequate form of testing because they lack sufficient realism. However, this is often not the case if the task and job analysis provide the necessary information for setting-up realistic standards and conditions, and for creating a device that tests the task in question. The major problem with this approach is that it takes a good deal of time, requires technical and educational resources, and is costly. However, if the device can be used in testing and training, it becomes an important addition to the soldier's job training.

Another way in which job stimuli can be presented is through the use of real equipment. Sometimes, the only way job stimuli can be presented is through the use of real equipment as is the case in laying a road bed or using a ship's boom to move cargo. Obviously, it would be inefficient to lay miles of road or tie up a large transport simply to test a soldier. This problem has been partially solved through the use of the performance certification component (PCC) described earlier.

While similar to the HOC, the PCC presents its own unique problems and will not be discussed here.

Stimuli Surrounding The Examinee.

The stimuli surrounding the examinee are the test conditions. The most important aspect of the test conditions in the HOC is that they be standardized for all examinees. Standardization helps insure that all examinees are treated essentially the same and the performance situation is defined clearly. Standardization is achieved by issuing instructions that specify what conditions must be met before the test is administered. The instructions also supply the limits of the test conditions. Conditions are especially important when the test is administered outdoors. At a minimum, the following environmental conditions should be considered: light, precipitation, visibility temperature and noise. If the conditions do not fall within those specified by the test developers, the soldier receives a "Not Rated". Therefore, conditions must be sufficiently broad enought to permit testing under many conditions and limited enough to prevent one examinee from having an unfair advantage over another.

Response Aspects.

Response aspects can be discussed in terms of process or product scoring. In process scoring the behavior of the soldier is observed and scored. In product scoring, the result of the behavior is scored. We believe that it is better to use product scoring whenever possible. If this is done the scorer can take sufficient time to review the product and can get the opinion of other scores. Product scoring also provides a record of the performance which can be used to monitor the scorers.

A problem that is sometimes encountered in scoring deals with error tolerance. This is a problem when perception is minute as in the sighting of a weapon or in reading a measuring device that has minute calibrations. Scorers must be made aware of this problem and a realistic range set in order to reduce perceptual errors. Scorer fatigue and bordom is another problem that must be eliminated if accurate scores are to be obtained. Giving the scorers breaks from scoring and allowing them to score different parts of the test helps minimize this problem.

An Example

At this point, it may be informative to show a scoresheet used in a hands-on test. Figure 1 depicts the scoresheet for assembling and reparing a tactical FM radio. The scoresheet contains seven critical performance measures. The task is product scored. All performance measures are scored after the soldier has completed the task. The scorer does not score the performance measures by observing the soldier during the process of putting the radio together. In this case, correct process is determined by correct product. Essentially, the task can be scored GO or NO-GO with performance measures five, six, and seven.

The other performance measures are there for several reasons. First, soldiers use the schoresheets in their SQT notices as guides in preparing and training for SQT. Secondly, feedback is more specific and useful when these other performance measures are listed. Performance measure seven "did not damage battery" is used because during tryouts of this test, when performance measure seven was not yet included, many soldiers broke the batteries by incorrect assembly procedures. Thus, this performance measure was added to cover critical process elements of the task.

The reference to the notes in performance steps 6 and 7 refer to special scoring instructions at the bottom of the page. These notes provide more detailed information about scoring steps 6 and 7. The scoresheet also provides a place for recording any additional information needed to explain a NO GO. For example, if a soldier failed, the scorer would write down the reason for the NO GO.

PREPARE (AN/PRC-77) RADIO SCORESHEET (PREPARE/OPERATE TACTICAL FM RADIOS) (AN/PRC-77, AN/VRC-64, and AN/GRC-160) (071-11A-0930)

PERFORMANCE MEASURES (Product Scored)	PASS	FAIL
1. Assembled radio to include the antenna and antenna base.	____	____
2. Set the correct frequency	____	____
3. Turned radio on so that it can transmit	____	____
4. Turned volume up so that radio can receive	____	____
5. Completed the task in 2 minutes	____	____
6. Control station contacted (by scorer)	____	____
7. Did not damage battery (see note 2)	____	____

	GO	NO-GO
STANDARD: The soldier is scored GO if he passes all the performance measures. If he fails any performance measure, he is scored NO-GO. If the soldier gets a NO-GO, record on the scoresheet any additional information required to show the cause of the NO-GO.	☐	☐

ADDITIONAL REASON(S) FOR NO-GO:

____

____

____
Scorer's Signature

NOTE 1: At the end of 2 minutes when "STOP" is given, the scorer conducting the test will go around the semicircle to each radio and conduct a radio check with the control station monitored by the SA. If the station can be contacted, performance measure No. 4 is scored PASS. If the control station cannot be contacted, the scorer must determine if the radio was properly prepared for operation. If the scorer is able to place the radio into operation by correcting an error made by the examinee, performance measure No. 4 is scored FAIL. If the scorer determines that the radio is unserviceable through no fault of the examinee, the examinee will be retested.

NOTE 2: After the radio check, the scorer will physically check the radio battery to determine if the battery has been broken. If the battery is broken, the soldier will receive a NO-GO.

FIGURE 1

Administrative Considerations

When compared to routinely administered written tests, the hands-on test presents a considerable challenge to test administration. In order to reduce problems associated with administrative procedures, the hands-on test undergoes field tryouts during development. Test developers use tryouts to check the set-up of the test, the clarity of the instructions, the acceptability of the test to soldiers and to experts and the feasibility of the test.

A common problem encountered is the failure of equipment. In order to deal with the problem, all instructions say that the examinee has the option to stop the test at any point if a malfunction in the equipment is suspected. If there actually is a malfunction, the examinee is not penalized. However, if there is no malfunction, the examinee is penalized. While some may consider this requirement harsh, it nevertheless represents a realistic evaluation of performance, because deciding if equipment is functional is a part of every task. The scorer normally would not purposely program a fault into a piece of equipment, and makes every effort to insure that the equipment is functional. The only exceptions to this are in cases where the task requires the examinee to discover some fault in the equipment. Many tests make use of these so-called "troubleshooting" tasks. The only information the examinee has is that there is a flaw in the equipment. He/she of course, has no idea what flaw was programmed into the equipment. After the test is written and published, rehearsal is used so that administrators can become familiar with the test. The rehearsal is done as close as possible to the actual test date so that scorers and administrators will have recent experience with administration of the hands-on test. This also gives scorers and administrators time to solve unique problems and clarify instructions.

Another requirement, that of having scorers be of the same or a related MOS of those being tested, also helps reduce the possibility of scorer error. When the scorers have familiarity with the task, they are better able to identify and solve problems.

Another aspect of ITED's plan to reduce and solve testing problems is the use of a special hot line telephone number that test administrators can call to ask questions or get information about the test. However, even this is of limited use when the test is in progress or it takes time to find the answer to a problem. While much can be done to reduce the occurance of problems during a test there is, of course, no way that all problems can be foreseen and planned for in advance. For this reason, it has been a policy of ITED that no soldier should be penalized because of a problem that is beyond his control.

If some unforseen problems appears, such as equipment failure, equipment unavailability or weather change, the soldier receives a score of "not rated". A "not rated" does not help or hurt a soldier in his/her overall testing program. The "not rated" provides a useful option in situations that are beyond the control of examinees and administrators. In the Infantry tests only 1% of the soldiers received a "not rated". A statistic which may suggest that many of the field problems have been solved.

Another administrative problem is the manpower required for hands-on testing. This problem will probably never be solved completely to a commander's satisfaction because he/she must provide numerous personnel to administer and score the test. However, if the testing provides useful information and really enhances the effectiveness of training the manpower requirement becomes a much more tolerable problem.

RELIABILITY AND VALIDITY

Reliability and validity are beyond the scope of the paper, but it is important to mention how they are assessed. Perhaps, the most important aspect of reliability to consider is inter-rater or scorer reliability. Inter-rater reliability is assessed empirically and must be at least 80% before a scoresheet is accepted.

The other aspects of reliability, i.e., stability, equivalency, and homogenity, have not been considered because they are not germain to hands-on testing. Moreover, the fact that the hands-on test is a criterion-referenced test argues for not considering these aspects of reliability.

A more pertinent consideration is that of test validity. Content validity is established for all hands-on tests by having experts score a hands-on test. If 75% of these so-called experts agree that the task is an important part of their MOS and is a fair way to find out if a soldier can do the task, the test is deemed to have content validity. While not an empirical method of assessing validity the method provides logical validity and insures that the task is representative of an MOS. It also insures that soldiers who will take the test perceive it as a relevant and fair form of testing.

In conclusion, it should be noted that the Army is committed to hands-on testing and will work to solve the problems associated with performance testing.

AN OVERVIEW OF THE SKILLS QUALIFICATION TEST DEVELOPMENT WORKSHOP

October 1977

Military Testing Association Conference
San Antonio, Texas

Charlotte H. Campbell, J. Patrick Ford
and Roy C. Campbell

Human Resources Research Organization
Fort Knox, Kentucky 40121

OVERVIEW OF SQT DEVELOPMENT WORKSHOP

The Training Developments Institute (TDI), the Individual Training and Evaluation Directorate (ITED), and the Human Resources Research Organization (HumRRO), are conducting workshops that present the basic principles of developing criterion-referenced tests as the principles apply to developing a Skill Qualification Test (SQT). The workshops were developed by HumRRO under contract with the Army Research Institute (ARI).

In this paper I will describe the need for such a workshop, the constraints on the workshop, and the characteristics of the workshop.

NEED

The SQT is a highly complex system in form, development, and administration.

SQT is a criterion-referenced test; that is, performance on the test is measured against a standard determined in advance by an analysis of job performance requirements. Criterion-referenced testing, in various forms and under various names, has been around for a long time. But the SQT is, in form, development, and administration, different from other forms of criterion-referenced systems. The need for training, as provided by the workshop, arises from these differences and the problems associated with them.

In form, the principal difference is in the three components, or types of test mode, within an SQT. An SQT consists of:

Hands-On Component (HOC)

Performance Certification Component (PCC)

Written Component (WC)

The Hands-On Component represents the most common type of criterion-referenced testing. The examinee performs a task, or part of a task, under standardized conditions, and is evaluated against a standard of job performance. Even though hands-on testing is most clearly criterion-referenced, an SQT may have no hands-on component.

The Performance Certification Component also resembles traditional criterion-referenced testing. The conditions under which the test is performed are not always standardized, but the allowable range of conditions is specified. A soldier's performance is usually evaluated by a

supervisor as he performs the task on the job. The PCC is clearly criterion-referenced: performance is measured against a standard. But some SQT will have no PCC.

The third component of the SQT is the Written Component. In appearance, it resembles any other written test: there are item stems and alternative responses, and the examinee chooses the right response or responses. But where most written tests sample a domain of knowledges, the Written Component is criterion-referenced: the examinee performs portions of tasks, or indicates how the task is performed. The scorable units, of 2-10 items each, cover discrete tasks, and the item alternatives represent the actual alternatives that the soldier encounters on the job. The examinee then receives a score, GO or NO GO, for each task. All SQT will have a written component.

Thus the form of an SQT is complex. Problems arise in selecting tasks to be tested, adapting task analysis data to make them suitable for constructing of criterion-referenced tests, and determining which of the three components is best suited for testing each task. "How we used to do it" sometimes interferes with how it must be done.

In addition to form of SQT, problems also arise in development. Development of the SQT involves much more than the construction of the test for each task. The primary problem in development arises from the requirement that an SQT must be validated. Unlike most criterion-referenced testing systems, an SQT must be tried out for reliability and validity before it can be administered in the field. For the HOC, the procedure involves checks of scorer reliability and test feasibility; for the PCC, scorer reliability, test feasibility, and systematic monitoring of testing must be ensured; for the WC, predictive and content validity are of concern. The procedures for validating an SQT are unique.

In the area of administration, the SQT also produces some distinctive requirements and problems. Everyone in a particular MOS skill level, worldwide, will take the SQT during the same test period. Besides being large scale, testing is also decentralized. Test Control Officers at each installation will conduct SQTs. For the HOC and PCC, this means that developers must prepare very precise performance measures, test conditions, and instructions to the Test Control Officers, scorers, and examinees. For the WC, the large scale, decentralized testing means that every response must wind up as a mark on a machine-scored answer sheet.

These unique characteristics of an SQT, in form, development, and administration, have created considerable problems. Even experienced test developers, even experienced criterion-referenced test developers, have found that developing an SQT is not an easy job. And at most Test Development Agencies (TDA), the SQT developers are not test experts, but subject matter experts. Their expertise is vital in SQT development, but not sufficient.

Early in the history of SQT development, ITED became aware of recurring problems on nine major tasks or aspects of the test developer's job:

Select Tasks for Testing

Review Task Analysis

Allocate Tasks to Components

Construct Hands-On Component

Tryout Hands-On Component

Construct Written Component

Validate Written Component

Construct Performance Certification Component

Prepare SQT Notice

Guidance on the nine tasks was published as the Guidelines for Development of Skill Qualification Tests. In addition to this document, however, a need was perceived for a controlled, systematic approach to training and assisting individual developers in the implementation of the principles contained in the Guidelines. The SQT Development Workshop was proposed as a means to provide monitored practice in the skills involved in the nine tasks.

The overall objective for the Workshop is to prepare people at Test Development Agencies to perform these nine tasks, and to apply them to their own SQT development.

CONSTRAINTS

The workshop had to accommodate three constraints. The first of these is that it had to be exportable. While responsibility for development of the course was assumed by the U.S. Army Training and Doctrine Command (TRADOC), ultimately the implementation of the course is the responsibility of the individual TDA. TDA traditionally experience considerable turnover among SQT personnel. TDA must be able to repeat the course as often as their needs dictate.

The decision was made to make the course a part of the total Faculty Development Program under the direction of TDI. At the TDA or school level, the course would be the responsibility of the Staff and Faculty

Development section. The requirement then, was that the course be exportable to the extent that it could be taught to staff and faculty personnel who would then act as course managers at their TDA and conduct the course as needed to meet their own needs.

The second requirement was that the course be self-paced. While the TRADOC training philosophy incorporates self-pacing in its instructional model, this was not the only basis for this requirement. Persons who are assigned to SQT development have a variety of experience. Some know a great deal about testing but little about the practical limitations of SQT. When participants come to the workshop, their actual experience with SQT ranges from absolutely no prior exposure to SQT to two years working in SQT Development. Thus, as the need to learn about various aspects of SQT varies, the workshop had to allow individuals to work at their own pace.

The third constraint related to the time of the workshop. You may have heard that training should be limited only by the amount of time required for students to master the objectives. You may even have said it. But there is almost always an outside limit. For this workshop the limit is two weeks. Managers are just not willing to allow people to be away from their desks for more than ten days to learn to develop an SQT.

These constraints have been faced by other developers. As part of the total Faculty Development Program, TDI has successfully implemented a Criterion-Referenced Instruction (CRI) Workshop, developed by Mager Associates. This CRI workshop has become the basic foundation for a family of staff and faculty development programs which will provide the necessary in-house training capability in each TRADOC training facility.

According to the CRI model, the overall objective for a training program is broken down into subordinate objectives, and training is presented in modules corresponding to these subordinate objectives. Within some limits, participants choose the sequence in which they will tackle the modules. At the beginning of each module, the objective for that module is stated, the criterion test is described, and resource references for the material are listed. Each participant decides individually how much he must study and practice to pass the criterion test. A course manager monitors student progress, evaluates criterion tests, and serves as a learning resource when required by the student.

This basic framework was followed for development of the SQT Workshop.

CHARACTERISTICS OF THE WORKSHOP

The workshop is designed for worker-level development personnel, that is, the individual who actually must produce an SQT. Because of the detail in which the material is presented, it is not intended for senior or most middle management level personnel.

The workshop objective, to prepare people to apply the principles in the nine tasks, was broken out into 34 subordinate objectives. For example, one such subordinate objective (within the task, "Construct Hands-On Component") is: "Using a task analysis for a task allocated to the HOC, construct performance measures for process scoring, product scoring, or combination scoring." Thirty-four modules were prepared for these subordinate objectives. Each module contains explanatory text, examples and practical exercises. The examples and practical exercises are for the most part based on common military tasks. Sample tasks were chosen to illustrate the principles being discussed, and are intended to be familiar to most course participants.

For each module, there is also a criterion test. Each criterion test involves one of two types of material. In some tests, the material is standard, that is, the participant is given a situation, task, or other information, and applies the concepts put forth in the module to satisfy the requirements of the test. In others, the participant is expected to work with a task and material of his choosing from the MOS and skill level that he will be working with during SQT development.

The balance between the two types of material was not easy to achieve. Standard tests are amenable to very specific feedback, and make the role of the course manager easier. However, requiring the developer to use his own material helps to overcome the "My MOS is different" syndrome by showing developers the adaptability of the course materials. In this way, the participant develops a greater appreciation of the flexibility and relevance of the principles to his own job. Approximately one-half of the criterion tests involve developers in using their own tasks.

In the workshop, the nine major tasks discussed earlier were grouped into seven phases of skill development. These seven phases are necessary for complete development. Although emphasis in the workshop is on the individual modules, not the phases, in the time remaining I will briefly outline what is involved in each of the seven phases. (See Figure 1.)

The first phase, for all participants, is the analysis and planning phase. At the beginning of the workshop, participants select an MOS with which to work, one with which they are familiar. They begin with ten tasks from one skill level of that MOS. In one module, participants identify sources of information on each task that are objective indicators

of need for evaluation. In another module, participants group the ten tasks according to the extent of known performance deficiencies. These modules lead participants to select for testing those tasks which promise the greatest payoff in testing. From the ten tasks, a course manager then selects five tasks with which the participant continues to work. Then, in the criterion test for the task analysis module, participants review and, if necessary, revise existing task analysis data for those five tasks to make them suitable for test construction. The final module in the analysis and planning phase covers allocating tasks to components. In the criterion test, participants assign each of their five tasks to the HOC, the PCC, or the WC. High skill physical tasks are allocated to the HOC or the PCC; mental tasks and low skill physical tasks are allocated to the WC.

After participants finish the analysis and planning modules, they branch into either the HOC construction phase or the WC construction phase. During the construction phases, participants work with the tasks selected earlier. For the HOC construction, there are modules for some preliminary decisions called for in the Guidelines. Then they work on modules which require that they construct two complete hands-on scorable units, to include performance measures, conditions, examinee instructions and scorer instructions.

The WC construction phase also requires participants to write scorable units for tasks they selected. They practice constructing two kinds of written test: written performance tests, which require examinees to perform part or all of a task, and performance-based tests, which require examinees to answer questions about how a task is performed.

After participants finish the construction phase for a component, they move to the validation phase for that component. Here, the activities and criterion tests are standardized, and address the analysis of data and revision of scorable units based on validation results.

The HOC validation procedure checks interrater reliability, acceptability, and feasibility. The modules cover locating faults based on a tryout with experts, computing scorer agreement, checking feasibility of a scorable unit, constructing a station-load table, and revising hands-on scorable units.

The WC validation procedure checks discriminant validity and acceptability. Three options for validation are available, based primarily on the number and types of soldiers to which the developer has access. The validation modules cover collecting self-ratings, locating faults based on a tryout with experts, validating written scorable units against hands-on tests, selecting a validation option and analyzing data on each of the three options, and revising written scorable units.

The revision modules cap each validation phase. ITED's policy in regard to validation is that the results give a basis for locating and correcting faults in a test. The modules present troubleshooting charts for hands-on or written tests. The activities and the criterion tests present summaries of results that indicate malfunctions on given items or performance measures. Using the troubleshooting charts, participants then modify the test to correct the probable causes of the malfunction. These modules call for additional practice of the skills acquired during the construction phases.

The sixth phase, dealing with the PCC, focuses not only on the procedures for constructing the PCC but also on procedures for validating and monitoring it. Participants again work with one of their own tasks. They describe how the test will be conducted, how it will be validated, what kinds of results would indicate units for follow-up checks, and how the checks will be conducted.

In the final phase, after participants have developed a scorable unit for each component, they prepare an SQT Notice. This is primarily a check on their mastery of the format for the Notice. It also provides a neat wrap-up of the course.

In this way, workshop participants work through a full cycle of SQT development in about ten days. Workshop materials will be revised based on our experience with 15 TDA currently being trained. Reaction from those who have already received the training has been overwhelmingly favorable. Even individuals who have had no previous contact with test development or with SQT have expressed confidence in their ability to fit into the SQT system after taking the workshop. Likewise, participants, who have had prior work with SQT state that the workshop has improved their skills and capabilities. The revised workshop will then be added to the Staff and Faculty Development courses at the TDA. The concentrated practical work in this workshop will thus be part of the on-going TRADOC support of TDA to accomplish the unique goals of SQT.

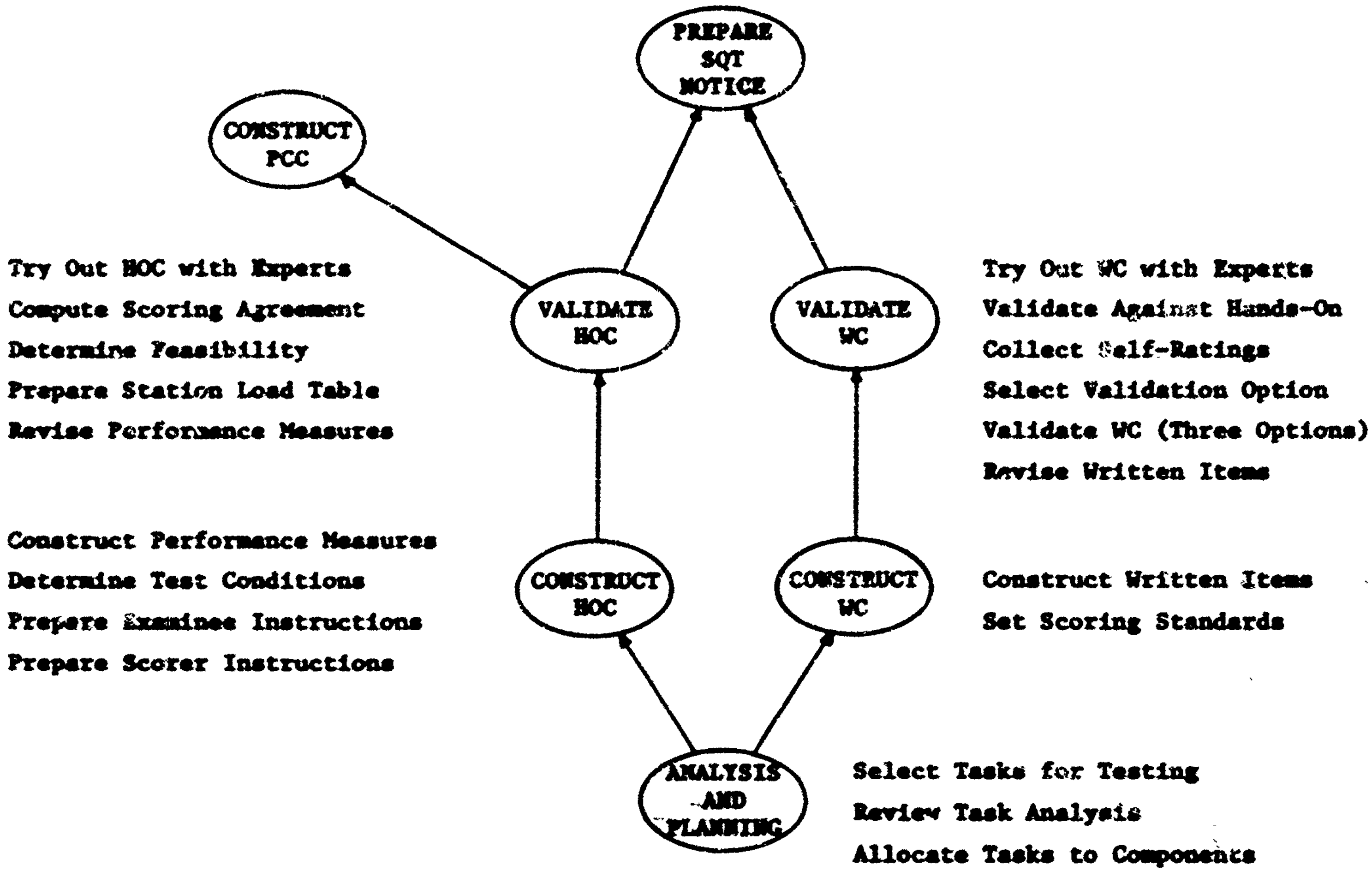


FIGURE 1. Phases of SQT Development Workshop

TAXONOMY OF TERMS IN JOB ANALYSIS

Major J. L. Mitchell, USAF
USAF Occupational Measurement Center
Lackland AFB, Texas

The term "Job Analysis" is used to refer to the whole class of studies of occupational information as well as to the study of a single job. The same language is used to discuss job data used for a variety of purposes, from the engineering of a new cockpit design to the breakdown of tasks into elements which can be used to develop criterion objectives for training. Such a diversity of usage for the term "Job Analysis" creates some difficulties in communication and sometimes results in evaluating job analysis systems which are designed for different purposes and have radically different procedures. This is, of course, absurd and a waste of valuable research time. To preclude such misunderstanding and waste, we need to develop and consistently use a more refined set of categories which will more precisely communicate the type of job analysis and its purpose.

Obviously, as an initial step we must begin by defining just what is meant by the generic term "Job Analysis". McCormick and Tiffin (1974) have defined it in the following way:

> "Job analysis can be considered as embracing the collection and analysis of any type of job-related information, by any method, for any purpose. Perhaps it can be defined more generally as the study of human work (McCormick & Tiffin 1974:49)."

As you can readily see, this very general definition is so broad as to include any possible type of study involving human work activities. This would encompass the entire spectrum of occupational studies; from the vaguely worded narrative job description (which has been particularly notable in its use to describe higher level executive and management positions) to the most percisely specified and quantified job analysis systems used for the more repetitive or consistent positions (those which can be done by checklist).

A number of reviews and critiques have lamented the state of Job Analysis in this country. Kershner (1955) observed, "As is patently evident, job analysis has been a sort of handmaiden serving in various ways a variety of needs and all the while floundering in a morass of semantic confusion." This is a bleak picture indeed, but is probably an accurate picture of the state of Job Analysis at the time Kershner was writing in 1955.

Prien and Ronan in a 1971 review of Job Analysis suggested that a considerable amount of progress had been made, particularly in the military services and in some of the more quantified job analysis systems such as that developed by McCormick in the Occupational Research Center at Purdue University. While they cited six major research areas which remained to be resolved, Prien and Ronan seem to believe that a credible amount of progress had been made in the decade and a half since Kershner's critical comments

In a more recent review, McCormick examined the area of Job and Task Analysis for the new Handbook of Industrial and Organizational Psychology. In evaluating the overall area, McCormick concluded that,

"... with some notable exceptions, the study of human work has generally been more in the domain of the arts than of the sciences. Perhaps to express it differently, the study of human work (which occupies a major part of man's lifetime) probably has not generally benefited from the systematic, scientific approaches that have been characteristic of other domains of inquiry, such as the study of physical phenomena, biological phenomena, or of the behavior of man himself (as through psychological and sociological research)." (McCormick 1976:654)

The few "bright spots" which McCormick sees in this whole area include the work done by the US Training and Employment Service (UST&ES) in occupational classification, the work of task analysis done in the military services (and particularly the development of CODAP by the Air Force Human Resources Lab), and the few civilian systems which focus on the quantification of job information (such as McCormick's own Position Analysis Questionnaire system).

At the risk of being accused of bias, I would assert that the occupational research program of the military services is perhaps the brightest of the "bright spots" which McCormick outlined. Of course it is impossible to evaluate these three systems or approaches since they are designed for different purposes, use entirely different procedures, and serve different populations. However, I think it would be accurate to say that the CODAP-based military job analysis systems have probably had the greatest impact in the sense of terms of providing a data base and systematic analysis of that data base to assist managers in making decisions concerning how jobs are structured, how they should be organized and defined, and what training is both necessary and relevant.

As early as 1959, Dr Carroll L. Shartle, who is perhaps the "grandfather" of occupational studies in this country, wrote that,

"Occupational information in the armed services has continued to develop and today the Department of Defense has one of the largest programs in developing and applying occupational information in the world. Technological changes make it necessary that occupational research be continued vigorously (Shartle 1959:8)."

Shartle also developed a basic taxonomy of terms used in the area of Job Analysis. Recognizing that our normal terminology dealing with jobs is often very loosely used in practice, he set forth certain definitions of terms in order to establish a more realistic framework for our study of occupations. His definitions are as follow:

Career. A career covers a sequence of positions, jobs, or occupations that one person engages in during his working life.

Occupation. An occupation is a group of similar jobs found in several establishments.

Job. A job is a group of similar positions in a single plant, business establishment, educational institution, or other organization. There may be one or many persons employed in the same job.

Position. A position is a group of tasks performed by one person. There are as many positions as there are workers in the organization.

Positions could be further broken down into tasks (those specific activities which taken together, make up the position) and elements (those very specific actions which comprise a task).

These terms, by and large, provide a comprehensive taxonomy. They are fairly consistently used in the military services, with the exception that we tend to collapse the first two categories into one when we talk of Military Occupational Specialties (MOS) or, in the Air Force, of Career Fields. Thus we tend to use the terms "career field", "specialty", or "occupational area" interchangeably to refer to what Shartle defines as Occupations and Careers. We should, perhaps, be more consistent and use his terms in order to differentiate between groupings of related jobs (an Occupation) and what an individual does during his time in the military service (a Career).

While this taxonomy of terms appears simple and straightforward, we have already seen that there is some degree of confusion in practical use. One of the problems is that we use the term "Job Analysis" to refer generically to the entire area of occupational information. At the same time, we also use it to refer specifically to the analysis of a group of related positions within a given work context or organization. In the Air Force, we tend to use the term "Job Description" regardless of the level of groups we are talking about. With CODAP we can generate quantative "Job Descriptions" for an individual, for groups of related positions (where is would be most appropriate), and even across groups to cover entire occupational areas. We would be more percise if we used Position Description for describing tasks involved in a unique individual's work, Job Description for outlining tasks for groups of related positions, and Occupation Description for summarizing the tasks involved across related job groups. This useage would let us grasp much more quickly the exact level of our descriptions, and could more percisely communicate the level of our analysis.

A very serious compounding factor, however, is the use to which the analysis is to be put. There is probably more variance in procedures, type of analysis, and meaning associated with what we plan to do with the data than there is with the level of grouping. A task, job, or occupational analysis for the purpose of engineering design of a new weapons system cockpit is necessarily considerably more detailed than would be required if the analysis is to be used only for examining the job classification system. And yet, at present, we continue to use the terms Job, Task, or Occupational Analysis as if they are equivalent regardless of the purpose of the research.

We need to look, at least in a very general way, at the various possible uses of Job Analysis data. McCormick and Tiffin (1974) have published a summary table which very concisely summarizes some of the uses of the type of information which is normally obtained through job and task analysis. Their summary with some modifications, is shown as Table 1. As you can see from this display, the possible uses of job information are quite varied and implicitly demand quite different kinds of information. The kinds of job information needed for making personnel selection and placement decisions must necessarily be considerably different than would be needed for equipment design, although obviously they are to a degree related or interactive. The kinds of people available for selection does imapct on how you can design the equipment to be used on a job and the reverse is also true.

While we could discuss these various uses in considerable detail, it is more worthwhile here to simply note them and to understand that each possible use of job information has its own unique requirements. While they may be related to one another to a small or large degree, the information needed is by no means identical. The same is true of the levels discussed earlier. The types of information needed to adequately describe a specific position are not necessarily what is needed to properly characterize an occupation. The former by necessity must be much more specific and detailed than the latter. Further, what can be done with the data is also relevant. It is useless to assess the similarity of a position with itself; obviously it is identical. However, such a contrast in terms of similarity is just what is needed when we study job groups or when we wish to assess the degree of relationship between various occupational groups or career patterns. Thus, both the level and the purpose of the analysis are relevant and we need some way to communicate both of these in our taxonomy of terms.

Before proceeding to an obvious solution to these taxonomic problems, it might be worthwhile to cite at least one instance where this type of basic definitional difference has resulted in a serious scientific problem. Recently, a draft paper by a Navy research contractor has been circulating among the various military organizations involved with occupational analysis. The main point of this paper was to compare the relative cost effectiveness of the contractor's method of determining Navy course content requirements and that of the CODAP-based USAF Occupational Survey Program. This researcher asserted that he could achieve the same objective of establishing what needs to be trained by using a job analysis expert and a small group of training specialists qualified in the given occupation without needing to do an expensive survey or use a sophisticated computer program.

TABLE 1

USES OF JOB AND TASK ANALYSIS DATA

WORK AND EQUIPMENT DESIGN	ADMINISTRATIVE CONTROL	PERSONNEL ADMINISTRATION	OTHER USES
Engineering design	Organizational planning	Personnel recruitment, selection, & placement	Planning educational curricula
Methods design	Manpower planning and control	Training & personnel development	Vocational counseling
Job design		Performance measurement & rating	Job classification systems
		Wage and salary administration	
		Labor relations	

(Adapted from McCormick & Tiffin 1974: Fig. 3.1; page 49)

This report could be challenged on a number of points. For example, the "improved" criteria of success is simply the opinion of an analyst and a trainer, which is hardly an objective criterion. Further the contractor asserts that no one has been concerned with the question of the validity of survey data; in this he largely ignores the concerns expressed by Kershner (1955), by Prien and Ronan (1971), by Ash & Kroeker (1975), and by McCormick (1976), as well as others.

Quite aside from such criticisms of what the researcher said, a much more basic issue is involved when his entire approach attempts to compare the results of a training-specific task analysis with what is possible with an Air Force occupational analysis, where the objective is much broader in that career structure, classification, promotion testing, and other issues as well as training are involved. He biases his results by essentially saying that by focusing on one particular objective, he can accomplish it more economically than is possible with a more comprehensive system. In fact, he has not proven even that even though he ignores the other potential uses of job information. The more sophisticated CODAP-based system can address the question of whether training should be given at all which is not feasible in the proposed training expert system. The contractor ignores the fact that changes made in the classification system also have a direct impact on training requirements.

The contractor's system also makes the assumption that management has already made most of the necessary decisions as to what the training population will be and the degree of proficiency required, and related issues. He completely misses the point that the USAF occupational survey program is designed primarily to provide Air Force management with the information to make data-based decisions about an occupational area. As a spin off from gathering such information, data is also available to be used in training design and decisions. Additionally, within the Air Force system, a procedure much like that proposed is used when decisions are to be made about specific training content, except that in the Air Force program we are currently having technical representatives from using commands (those who use our technical training graduates) assist in this type of decision making ("Scrubdown" or "Technical Training Systems Review" Projects).

Finally, the Navy contractor asserts that decisions made in his system by an expert and the trainers should be used as the basis for designing occupational structures - that is, that training should drive the classification system. This, of course, goes much beyond his limited data (which is largely anecdotal). The Air Force system is predicated on just the opposite assumption - it is necessary first to determine how people are or should be utilized using the most complete data base available (which includes occupational survey data, the expertise of subject-matter specialists, the experience of senior managers, and any other relevant data). Once the utilization pattern has been established, then the occupational structure and the training will follow. Rather than expecting management to make all the decisions needed to specify training programs, the Air Force occupational survey system is designed to assist those managers by providing a data base reflecting current utilization for use as a starting point in the decision process.

This brief review of a contractor's research is cited here as perhaps a rather extreme example of attempting to compare a "job analysis" for a limited purpose (training) with a "job analysis" designed for a higher level of generality and for various purposes. If we used more percise definitions, it is evident that he attempts to contrast a task analysis for training with an occupational analysis which is multipurpose (classification, structure, training, etc.). Stated in these more precise terms, his attempt is both unrealistic and unscientific. It does, however, serve as an excellent example of why we need to be more explicit and consistent in our use of Job Analysis terminology.

One obvious solution for this need for more percise language is to use a composite title which would include both the level and the purpose of the analysis. Table 2 displays a matrix of the levels of analysis as well as some of the uses of job analysis. A few of the cells in this table have been filled in to illustrate the type of terminology which could be used to very concisely depict the type of analysis being undertaken. Thus, a task analysis being done specifically to determine training requirements would be a "Training Task Analysis" while a task analysis being done for the purpose of human engineering would be an "Engineering Task Analysis". This system would be most useful at the more specific levels and in those cases where only one purpose is to be met.

For the higher levels of analysis (particularly in the occupational analysis or career analysis categories), it may be best to drop the specific purpose designation, especially where there is more than one purpose to be served. Thus, where no qualifier is included in the title, we could assume a multipurpose study. Where only a single purpose is undertaken, such as the study of job satisfaction then a more specific title would still be appropriate (a Career Satisfaction Analysis? etc.).

While a more flexible and refined taxonomy of terms will certainly not solve many of the complex and difficult problems of the study of human work, it is a starting point for more percise communication about the kinds of analysis we do. As such, it represents a better way of doing business and is therefore proposed for your consideration and use. Such a system has much to commend its use and should provide a foundation on which we can build a more comprehensive dialogue among ourselves and with the managers whose need for data-based decisions we are in the business to serve.

TABLE 2

MATRIX OF LEVELS AND USES OF JOB INFORMATION WITH SAMPLE CELL ENTRIES

USES/LEVELS	ELEMENT	TASK	POSITION	JOB	OCCUPATION	CAREER
Engineering design	Engineering element analysis	Engineering task analysis				
Methods design				Job methods analysis		
Job design			Position design analysis		Occupational structure analysis	
Organizational planning						
Manpower planning			Manpower position analysis			Career manpower analysis
Personnel selection & placement				Job selection analysis		
Training		Training task analysis	Position trng analysis		Occupation trng analysis	Career trng ana
Performance msmt.						
Wage & salary				Compensation job analysis		
Curriculum dev.						
Counseling		Task counseling ana.		Job counseling analysis	Occupational counseling ana	Career counseling analysis
Etc... etc...						

REFERENCES CITED

Ash, P., and L. P. Kroeker Personnel selection, classification, and placement in Rosenzweig, M. R. & Porter, L. W. (Ed), Annual Review of Psychology, 1975, 26, Palo Alto, CA: Annual Reviews, Inc.

Kershner, Alan M. A Report on Job Analysis, ONR Report ACR-5, Office of Naval Research, Department of the Navy, Washington DC, 1955.

McCormick, Ernest J. Job and Task Analysis. Chapter 15 in M.D. Dunnette (ed), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally College Publishing Company, 1976; 651-696.

McCormick, Ernest J, and Joseph Tiffin. Industrial Psychology. Prentice-Hall, Inc.; Englewood Cliffs, N.J. 1974 (6th Ed).

Prien E. P., and W. W. Ronan. Job analysis: a review of research findings. Personnel Psychology (1971), 24:371-396.

Shartle, C. L. Occupational Information. Englewood Cliffs, N. J.: Prentice Hall, Inc., 1959 (3rd Ed.).

REFERENCE NOTE

The draft Navy contract research report mentioned in this paper has not been approved for public release by NPRDC, San Diego. At the request of the NPRDC contract monitor, the researcher's name has been deleted.

US NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM - A TEN YEAR PERSPECTIVE

Maurice D. Callahan

INTRODUCTION

SOME YEARS AFTER HE LEFT THE WHITE HOUSE, PRESIDENT TRUMAN IN REPLY TO A QUESTION, STATED THAT HE WAS THE OLDEST LIVING EX-PRESIDENT AND HE DESIRED TO REMAIN THAT WAY AS LONG AS POSSIBLE. AS A CHARTER MEMBER OF THE MTA, MY FEELINGS PARALLEL HIS. I WOULD LIKE TO REMAIN A LIVING CHARTER MEMBER AS LONG AS POSSIBLE.

THERE IS ANOTHER ORGANIZATION TO WHICH THIS PARALLEL APPLIES - THE PEARL HARBOR SURVIVORS ASSOCIATION - IT IS THE ONLY OUTFIT THAT I BELONG TO THAT DOES NOT HAVE ANY INPUT FROM THE BOTTOM - AND THE MEMBERS AT THE TOP ARE GRADUALLY DISAPPEARING. YOU MIGHT SAY WE HAVE NO UPWARD MOBILITY PROGRAM, OR PARADOXICALLY WE HAVE EMBRACED THE UP OR OUT POLICY TO ITS LOGICAL, ULTIMATE RESTING PLACE.

SO I AM HAPPY TO NOTE THE YOUNGER MEMBERS OF THE MTA, ESPECIALLY THOSE WITH THAT EAGER LOOK ABOUT THEM. AT TIMES THAT EAGER LOOK HAS TROUBLED ME - I AM NEVER QUITE CERTAIN IF THEY ARE EAGER TO AVOID THE MISTAKES WE MADE, OR EAGER TO REPEAT THEM. PERHAPS I MAY BE ABLE TO HELP SOME OF YOU AVOID SOME MISTAKES OR ONLY PARTIALLY RE-INVENT THE WHEEL.

SLIDE 1 ON

I WILL FOLLOW THE ABSTRACT AS SUBMITTED - HOW AND WHEN NOTAP STARTED, CONDITIONS EXISTING AT THE TIME - ASSUMPTIONS AND CONCLUSIONS MADE IN THE EARLY DAYS, BOTH ACCURATE AND INACCURATE AS PROVED BY SUBSEQUENT OPERATIONS, PRESENT STATUS OF THE PROGRAM AND A LOOK AT POSSIBLE FUTURE USES OF OCCUPATIONAL DATA.

SLIDE 1 OFF

SLIDE 2 ON

NOTAP WAS CONCEIVED BY THE SECNAV TASK FORCE ON RETENTION WHICH REPORTED OUT IN 1966. THE EXACT RECOMMENDATION WAS AS INDICATED ON THIS SLIDE.

SLIDE 2 OFF

SLIDE 3 ON

THE HISTORY OF THE PROGRAM IS AS FOLLOWS: (SPEAK TO SLIDE)

SLIDE 3 OFF

WHAT WERE THE CONDITIONS EXISTING IN THE LATE SIXTIES AND EARLY SEVENTIES? THE NAVY'S UNIQUE MANPOWER CHARACTERISTICS WERE AS LISTED ON THIS SLIDE.

SLIDE 4 ON

WE HAD AN EXPANDING NAVY THEN A CONTRACTING NAVY. IN THE EARLY SEVENTIES THE WASHINGTON PAPERS WERE CARRYING ARTICLES ABOUT MOVING OUT 25% OF FEDERAL GOVERNMENT PERSONNEL BASED THERE.

IT WAS NOT THE TIME AND PLACE FOR A NEW PROGRAM TO GO OPERATIONAL IN THE WASHINGTON AREA BUT, OBVIOUSLY WE MADE IT.

SLIDE 4 OFF

SLIDE 5 ON

TO REFRESH YOUR MEMORY, THIS SLIDE IS A COMPOSITE OF ALL THE INFORMATION WE COLLECT. (SPEAK TO SLIDE)

SLIDE 5 OFF

WHAT WERE SOME OF OUR EARLY RECOMMENDATIONS AND CONCLUSIONS? SOME ITEMS:

1. OBSERVATION/INTERVIEW. PRIOR TO CONSTRUCTING A TASK INVENTORY THE TEAM ASSIGNED INTERVIEWS A SMALL BUT HIGHLY REPRESENTATIVE SAMPLE OF PERSONNEL IN THE RATING - ANYWHERE FROM 40 TO 150 INCUMBENTS. THIS COSTS MONEY, ESPECIALLY TRAVEL FUNDS. IN ADDITION TO PROVIDING TRAINING FOR OUR TEAM, IT HELPS PIN-POINT PROBLEM AREAS AND PROVIDES PEOPLE TO PEOPLE CONTACT. MOST IMPORTANT, IT INCREASES THE CREDIBILITY OF OUR DATA. A PRUDENT INVESTMENT - WE THOUGHT AT ONE TIME WE WOULD HAVE HAD A TOUGHER TIME JUSTIFYING THIS EXPENDITURE.

DATA COLLECTION. ORIGINALLY, WE THOUGHT THAT COLLECTING THE DATA BY MAIL-OUTS WOULD BE CHEAPER. THE USAGE RATE FOR MAIL-OUTS IS ABOUT 37%; FOR-ON SITE ADMINISTRATION BY OUR STAFF, IT IS 99% PLUS. CONSIDERING PRINTING, TRAVEL, POSTAGE AND USAGE RATE, OUR DETAILED COST STUDIES HAVE CLEARLY SHOWN THAT ON-SITE ADMINISTRATION IS CHEAPER AND FASTER.

SAMPLE SIZE. ORIGINALLY, OVER 50%. NOW BETWEEN 20-30 PERCENT. IN ADDITION TO RATINGS, THE NAVY HAS NAVY ENLISTED CLASSIFICATIONS (CALLED NEC'S) WHICH ARE MOSTLY A SECONDARY LEVEL OF IDENTIFICATION BUT SOMETIMES CAN BE A TERTIARY. SOME RATINGS HAVE OVER 40 SUCH CODES. WE FAILED TO ANTICIPATE THE PRESENT DEMAND FOR DATA BY THESE NEC'S. AS YOU KNOW, THE LARGER THE NUMBER OF SUB-GROUPS, THE LARGER THE SAMPLE SIZES. ON THE OTHER HAND, THE SMALLER THE SAMPLE THE LESS INTERFERENCE WITH THE OPERATING FORCES.

ADMINISTRATIVE - BY THIS I MEAN THE AMOUNT OF TIME CONSUMED IN MAKING ARRANGEMENTS FOR THE OBSERVATION AND INTERVIEW AND DATA COLLECTION. WE JUST DIDN'T ALLOW ENOUGH TIME FOR THIS IN OUR EARLY ESTIMATES. SHIPS AND SQUADRONS ARE MOBILE AND THE NUMBER OF SHIPS HAS DECREASED. THUS WE VISIT THE SAME SHIPS MORE OFTEN AND WE ARE VERY MINDFUL THAT THE PRIMARY MISSION OF THE OPERATING FORCES IS NOT COMPLETING QUESTIONNAIRES.

STAFF. WE ANTICIPATED A GRADUAL BUILD-UP IN OUR APPROVED OPERATIONAL ALLOWANCE, BUT SUDDENLY WE WENT FROM FAMINE TO FEAST - AS A RESULT OUR DATA COLLECTION CURVE LOOKS LIKE THIS:

SLIDE 6 ON AND OFF

THIS IS PRESENTED IN A DIFFERENT PERSPECTIVE IN THE NEXT TWO SLIDES, AND OF COURSE, WILL RESULT IN SOME SCHEDULING PROBLEMS IN THE FUTURE.

SLIDES 7 & 8 ON AND OFF

USE OF DATA

IN SOME AREAS THE USE OF THE DATA HAS EXCEEDED EXPECTATIONS, IN OTHERS, IT HAS LAGGED EXPECTATIONS.

FOR EXAMPLE, ABOUT FOUR YEARS AGO WE WERE SERIOUSLY CONSIDERING DELETING PHYSICAL DEMANDS FROM OUR TASK INVENTORIES. NOW THE REACTION IS - IS THAT ALL YOU HAVE? IN TWO RATINGS WE ARE CURRENTLY COLLECTING PHYSICAL DEMANDS DATA AT THE REQUEST OF PRDC, SAN DIEGO TO ASSIST IN A LONGITUDINAL PHYSICAL REQUIREMENTS RESEARCH PROJECT.

THE INITIAL REACTION TO JOB SATISFACTION/DISSATISFACTION WAS MUCH THE SAME. NOW THE PRINTOUTS ON JOB SATISFACTION ARE VERY MUCH IN DEMAND. PRDC, SAN DIEGO HAS TAKEN OUR DATA ON FOUR SAMPLE RATINGS AND SUBJECTED IT TO EXTENSIVE FACTOR, MULTIPLE REGRESSION AND CORRELATIONAL ANALYSES. BASED ON THEIR PRELIMINARY FINDINGS WE ARE MAKING SOME MODIFICATIONS TO THE JOB SATISFACTION PART OF OUR TASK INVENTORY.

THE USE OF OUR DATA IN VARIOUS RESEARCH PROJECTS HAS EXCEEDED EXPECTATIONS. THE OFFICE OF NAVAL RESEARCH AND THE VARIOUS SYSTEM COMMANDS FREQUENTLY REFER COMPANIES, WITH WHOM THEY HAVE CONTRACTS, TO US BECAUSE OF THE INFORMATION IN OUR DATA BANK.

THE TRAINING COMMAND IS THE LARGEST BULK USER OF OUR DATA. OUR RELATIONSHIP HAS BECOME SO INSTITUTIONALIZED THAT, AFTER COMPLETING CERTAIN INTERNAL CHECKS, WE IMMEDIATELY FORWARD A LARGE BOX OF PRINTOUTS TO THEM AS INDICATED ON THIS SLIDE.

SLIDE 9 ON SLIDE 9 OFF

Our management bureau, the Bureau of Naval Personnel tasks NOTAP for special studies when changes to the Enlisted Rating Structure are comtemplated.

To summarize, the data is being used, in varying degrees, on most personnel management functional areas. We, of course, would like to see it used more in certain areas. An observation - most perceived problems filter down from the top, not up from the bottom. Thus, it is difficult to convince a manager that he may have problems in the future and you have helpful information. He probably has too many other current problems and he secretly hopes what you describe may not come to pass, at least, not during his watch. Thus, the biggest advances in the use of the data has taken place at those times when a real need became apparent. We are grateful for the foresighted and strong support given to us by our Program Manager on the staff of the Chief, Bureau of Naval Personnel.

FUTURE USES OF DATA

IN ADDITION TO THE EXPANDING USE OF OCCUPATIONAL DATA BROUGHT ABOUT BY REFINEMENTS IN COLLECTING AND PRESENTING THE DATA, WHAT DOES THE FUTURE LOOK LIKE FOR NEW APPLICATIONS OF THE DATA?

SLIDE 10 ON

THERE ARE TWO AREAS THAT APPEAR PROMISING. THESE ARE:

1. OCCUPATIONAL HEALTH AND SAFETY.
2. LEGAL CASES CONCERNING SELECTION PROCEDURES AND JOB REQUIREMENTS.

THE OCCUPATIONAL HEALTH AND SAFETY ACT WAS PASSED IN 1970. IT CONCERNS HEALTH IN THE WORK PLACE. A FORTHCOMING EXECUTIVE ORDER WILL APPARENTLY INCLUDE UNIFORMED PERSONNEL, SPECIFICALLY, UNDER THIS ACT.

PERSONNEL OF THE OCCUPATIONAL AND PREVENTIVE MEDICINE DIVISION OF THE BUREAU OF MEDICINE AND SURGERY WERE INTERESTED IN WHAT NAVY RATINGS WERE EXPOSED TO CERTAIN SUSPECTED CANCER-CAUSING AGENTS.

ACCORDINGLY, THEY CONVENED A PANEL OF INDUSTRIAL HYGIENISTS AND OCCUPATIONAL HEALTH PHYSICIANS AND ISOLATED THOSE TASKS FROM NOTAP JOB INVENTORIES (RATING RELATED) THAT WERE RELATED TO CERTAIN SUSPECTED CAUSATIVE AGENTS. FROM THE COMPUTER PRINT-OUTS, THE PERCENT OF MEMBERS WHO PERFORMED EACH TASK WAS THEN EXTRACTED. THE SAME PROCEDURE WAS FOLLOWED FOR EQUIPMENT ITEMS,

FOLLOWING OUR BREAKOUT OF USE OR REPAIR AND USE AND REPAIR. LAST, AND ALSO FOLLOWING OUR BREAKOUT, THE DATA WERE BROKEN OUT BY JOB TITLES, WATCH DUTIES, COLLATERAL DUTIES AND PHYSICAL DEMANDS. REGRETFULLY, OUR RELATIVE TIME STATISTICS WERE OF LITTLE VALUE FOR THIS USE OF THE DATA.

IN THE FUTURE, WE SHALL PROBABLY INCLUDE SEVERAL SPECIFIC TASK STATEMENTS THAT WILL BE OF VALUE TO THE MEDICAL PEOPLE.

THIS IS A HIGHLY SENSITIVE AREA. FURTHER, JOB INCUMBENTS ARE OFTEN UNAWARE OF ANY LONG-RANGE HEALTH HAZARDS IN THEIR WORK AREA.

THE FULL AND EVENTUAL IMPACT OF THIS ACT, AS AMENDED, IS A MATTER OF CONJECTURE. IT WILL HAVE, HOWEVER, A TREMENDOUS EFFECT ON OUR ECONOMY AS HARMLESS. SUBSTITUTE MATERIALS ARE DISCOVERED AND MANUFACTURED.

WITH REGARD TO THE SECOND AREA - LEGAL CASES - SELECTION PROCEDURES INCLUDE APPLICATION FORMS, INTERVIEWS, AND REFERRALS AS WELL AS TESTS. SUCH CRITERIA MUST BE JOB RELATED; LIKEWISE, ANY JOB REQUIREMENT MUST ALSO BE JOB RELATED.

OF COURSE, WITH REGARDS TO THE PROVISIONS OF THE PRIVACY ACT, AND OTHER LEGAL CONSTRAINTS, RELEVANCE IS LESS RELEVANT! IT IS NOT ONLY THE CIVIL RIGHTS ACT OF 1964 AND SUBSEQUENT EXECUTIVE ORDERS THAT HAS RESULTED IN VIRTUALLY A NEW DIMENSION; IT IS ALL MINORITIES, WOMEN AND, CURRENTLY, THE AGED. FOR EXAMPLE, IF A WOMAN OR AN ELDERLY PERSON TAKES AND FAILS SOME SORT OF STRENGTH OR DEXTERITY TEST, AND SUBSEQUENTLY CHALLENGES

SAID TEST, MANAGEMENT MUST PROVE THE TEST IS REASONABLY RELATED TO SUCCESS ON THE JOB OR IT IS A REASONABLE JOB REQUIREMENT.

SLIDE 10 OFF

THE COURTS HAVE, IN THE PAST, BASED THEIR FINDINGS ON TEST'S VALIDITY ON THE FOLLOWING: SLIDE 11 ON

1. PROFESSIONALLY DEVELOPED
2. SKILL RELATED
3. CURRENCY
4. ESSENTIAL - SAFETY, EFFICIENCY, MORALE?
5. REASONABLE - BUSINESS RELATED PURPOSES?

THE RELATIONSHIP OF OCCUPATIONAL TASK ANALYSIS TO MOST OF THESE FACTORS IS OBVIOUS.

SLIDE 11 OFF

ANY JOB REQUIREMENTS MUST BE JOB RELATED IN THE SAME CONTEXT THAT TESTS ARE. A WORD OF CAUTION - JOB TASK ANALYSIS INDICATES WHAT WORKERS <u>ARE</u> DOING. IF MANAGEMENT CHANGES THE JOB CONTENT (WHAT WORKERS SHOULD BE DOING) ANY RESULTING CHANGES IN JOB REQUIREMENTS MUST BE VALIDATED ALSO. MANAGEMENT, IF THE JOB ANALYSIS SO INDICATES, SHOULD EMPHASIZE THE NEGATIVE ASPECTS IN COMPARISON WITH THE FIVE CRITERIA LISTED ON THE SLIDE (IN OTHER WORDS SHOW DAMAGES FROM THE EXISTING CONDITIONS) IN ADDITION TO VALIDATING THE CHANGES.

IN CONCLUSION, SEVERAL BROAD OBSERVATIONS. MANY OF YOU HERE ARE IN RESEARCH - IN THE DEVELOPMENTAL STAGE WE KNOW IF THE PROGRAM WENT OPERATIONAL WE WOULD LIVE WITH OUR MISTAKES. SO A GOOD QUESTION FOR A RESEARCHER TO ASK HIMSELF IS "IF I SHOULD BE RESPONSIBLE FOR OPERATING WHAT I RECOMMEND, WOULD I MAKE THE SAME RECOMMENDATION OR, AT LEAST, INVESTIGATE FURTHER?"

SECONDLY, ALTHOUGH THERE IS WIDESPREAD USE OF THE DATA, THE FUTURE IS STILL WIDE OPEN FOR INCREASED USE OF THE DATA, ESPECIALLY IN THOSE AREAS WHEREIN PERSONNEL MANAGEMENT HAS PROBLEMS.

NOTAP

NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM

SEC NAV RETENTION TASK FORCE

" . . . . THE NAVY HAS ESSENTIALLY NO OCCUPATIONAL DATA GATHERING PROGRAM. LIMITED OCCUPATIONAL INFORMATION IS OBTAINED FROM UPDATING THE QUALIFICATIONS FOR ADVANCEMENT IN RATING AND FROM SPECIAL STUDIES IN NEW WEAPONS SYSTEMS . . . . HOWEVER, A SYSTEMATIC CONTINUING PROGRAM IS NOT IN BEING."

HISTORY

• SECNAV TASK FORCE	FY 1966
• EXPLORATORY DEVELOPMENT RESEARCH	FY 1966
• ADVANCED DEVELOPMENT RESEARCH	FY 1971
• OPERATIONAL STATUS	FY 1974
• NODAC	FY 1976

1573

NAVY'S UNIQUE MANPOWER CHARACTERISTICS

(CIRCA 1967)

1. 3100 PLUS DIFFERENT BILLETS
2. GENERALIST CONCEPT OF NAVY RATING
3. SEA AND SHORE BILLETS
4. INTRA-SHIP TYPE AND INTER-SHIP TYPE DIFFERENCES
5. LARGE NUMBER OF PERSONNEL INVOLVED - OVER 550.000

KINDS OF DATA NOTAP COLLECTS

BILLET/BILLET INCUMBENT BACKGROUND INFORMATION

- 40 STANDARD ITEMS

TASK STATEMENTS – (RATING RELATED)

- 400 - 620 TASKS
  - TIME SPENT ON TASKS ON A 5 POINT SCALE
  - INVOLVEMENT IN TASK ON 4 POINT SCALE

EQUIPMENT

- 20 - 400 ITEMS
  - FREQUENCY OF USE ON A 4 POINT SCALE
  - FREQUENCY OF REPAIR ON A 5 POINT SCALE

PHYSICAL AND MENTAL JOB CHARACTERISTICS

- 30 - 60 ITEMS ON A 5 POINT SCALE

JOB SATISFACTION

- 20 - 40 ITEMS ON A 7 POINT SCALE

COLLATERAL DUTIES

- 20 - 40 DUTIES

SPECIAL DATA

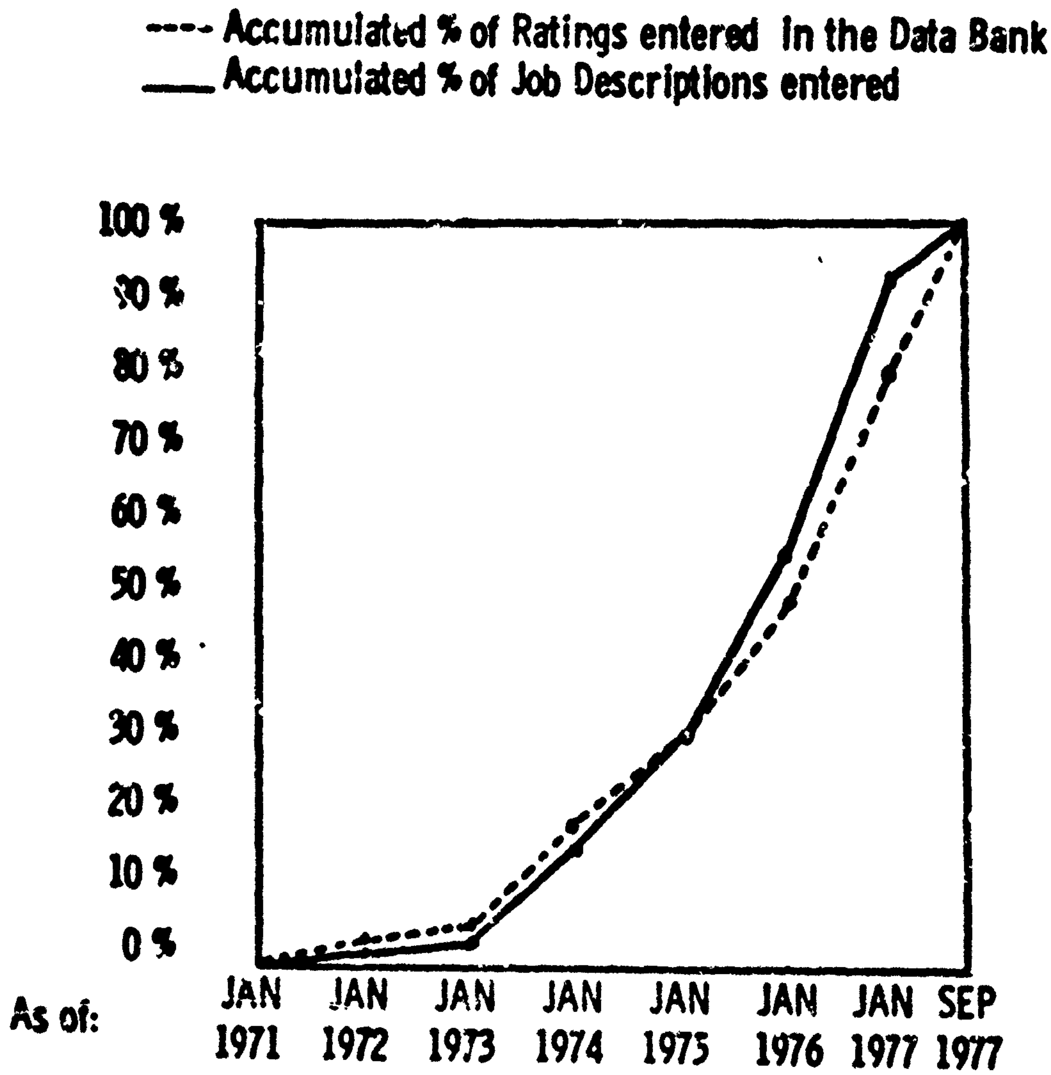
- 25 - 75 OPTIONAL CHOICES OF YES/NO OR TRUE/FALSE STATEMENTS

MISCELLANEOUS DATA

- 25 - 75 ITEMS OF ANY KIND OF BACKGROUND OR TASK-RELATED DATA

1612

PRODUCTIVITY CURVE



NAVY RATING

STATUS OF THE NOTAP DATA BANK

as of: 1 Sep 1977

Calendar Year	Number of Ratings Entered by Year	Percent of Ratings Entered by Year
1971	1	2 %
1972	2	3 %
1973	8	13 %
1974	7	11 %
1975	11	18 %
1976	20	32 %
Thru 1 Sep 77	13	21 %
	62	100 %

70 Ratings in the Navy

6 Small Ratings plus Hospital Corpsman and Dental Technician are scheduled or in process

JOB DESCRIPTION

STATUS OF THE NOTAP DATA BANK

as of: 1 Sep 1977

Calendar Year	Number of Job Descriptions Entered Each Year	Percent of Job Descriptions Entered Each Year
1971	727	1 %
1972	2,196	2 %
1973	10,970	12 %
1974	12,808	14 %
1975	23,108	25 %
1976	32,341	36 %
Thru 1 Sep 77	9,223	10 %
Total	91,373	100 %

STANDARD PACKAGE OF COMPUTER PRINTOUTS
FOR THE TRAINING COMMUNITY

Task Inventory Booklet & Optical Scan Booklet plus

PROGRAM		DESCRIPTION
NECSPG	-	NEC Listing by Paygrade
ACTCOD	-	Activity Code Listing
PRTDIC	-	Print Dictionary (Background Variables)
TITLES	-	Titles Listing (Duty & Task)
DIAGRM & GRPMBR	-	Diagram (Time) & Group Membership
PRTVAR	-	Print Variable (Background data)
DUVARS	-	Duty Variable (% of time spent by each member)
JOBDEC	-	Job Descriptions (Paygrade, Skill Level, Primary Diagram Stages, & Skill Levels within Stages, & Total)
GRPSUM	-	Summaries by % of Members Performing Each Task & % of Time Spent by All Members
ASFACT	-	Levels of Task Performance
AVALUE	-	Average Paygrade Performing Tasks
VARSUM	-	Variable Summary, (Worker Characteristics, Job Satisfaction, Watch Duties, Collateral Duties, Equipment Items, Etc.)

FUTURE USES OF TASK ANALYSIS DATA

1. OCCUPATIONAL HEALTH AND SAFETY
2. LEGAL CASES
    - A. SELECTION PROCEDURES
    - B. JOB (WORKER) REQUIREMENTS

COURTS / TEST VALIDITY

1. PROFESSIONALLY DEVELOPED ?
2. SKILL RELATED ?
3. CURRENCY ?
4. ESSENTIAL -- SAFETY -- EFFICIENCY -- MORALE ?
5. REASONABLE -- BUSINESS RELATED PURPOSE ?

AN ALTERNATE COMPUTER APPROACH TO THE ANALYSIS OF OCCUPATIONAL TASK-FACTOR DATA

Richard W. Dickinson
Occupational Research Program
Industrial Engineering Department
Texas A&M University

INTRODUCTION

The assumption is being made that the reader is at least vaguely familiar with the CODAP analysis system. While this paper is not intended to be a technical exposition on the manner in which the functions of the CODAP task-factor programs in question were duplicated, it is felt that some knowledge of CODAP is necessary to place the information presented here in the proper perspective.

Approximately a year and a half ago, the Occupational Research Program (ORP) at Texas A&M University was faced with the need to analyze and manipulate task-related information. Normally, this need would have been satisfied using the following task-oriented CODAP system programs:

COMGEN
FACSUM
FACSPC
FACSTD
PREFAC
TSKCAT
TSKCOR
TSKFAC

Unfortunately, the export version of the CODAP source codes does not contain these programs. This state of affairs lead ORP to search for other means of handling such data. Eventually, methods were developed for duplicating the functions of these eight programs through the use of SAS (Statistical Analysis System).

Description of SAS

SAS is a proprietary product of SAS Institute, Inc., a private company devoted to the maintenance and development of SAS. It is an integrated system of data management and statistical analysis conceived for use on I.B.M. or related equipment. This PL/1-like

language combines statistical versatility with extensive capability for data manipulation and report writing.

SAS can read data in almost any kind of format from any kind of file. For data transformations, SAS offers a complete library of mathmatical and statistical functions. The user can create new variables, delete observations and variables, accumulate totals, and execute statements conditionally. File-handling tools available in SAS include merging, sorting, interleaving, concatenating, subsetting, updating, and interactive editing of data sets. One of the most useful aspects of SAS is its ability to produce output from procedures in the form of SAS data sets. For example, SAS can put predicted values from a regression analysis in a data set for further analysis or manipulation.

Description of Task-Factor Programs

Below is a listing of the task-factor programs of interest along with a short description of their function. These eight programs were written to allow greater flexability in the manipulation, analysis and reporting of task-oriented data, which, up until this time, was inaccessable to existing CODAP programs.

COMGEN - Using user supplied program statements, this program performs mathmatical operations on factor data files.

FACSUM - Prints final report in specified sort format. This program can also create new files representing differences between various existing files.

FACSPC - Closely parallels program JOBDEC except that it applies to factor ratings instead of job descriptions.

FACSTD - Closely parallels program INPSTD except that it creates a rater history file instead of a worker history file.

PREFAC - Based on regression weights this program produces a deck of predicted score values.

TSKCAT - Produces a card deck with "1's" or "0's" representing specified tasks.

TSKCOR - Computes intercorrelation matrices and user-specified regression problems.

TSKFAC - This program is used to create new records on the job description file to be later referenced by other task-factor programs in the CODAP system.

Statement of the Problem

That task-oriented information can now be accessed similarly to that of information regarding jobs or persons adds significantly to the tools the job analyst can bring to bear on occupational problem areas. Unfortunately, the source code in which the eight task-factor programs were written has not been converted for use on non-Univac equipment. Such a situation effectively isolates users of CODAP with non-Univac equipment from benefiting from the convenience and flexibility these programs offer in the handling of task data. It was for this reason that ORP turned to SAS.

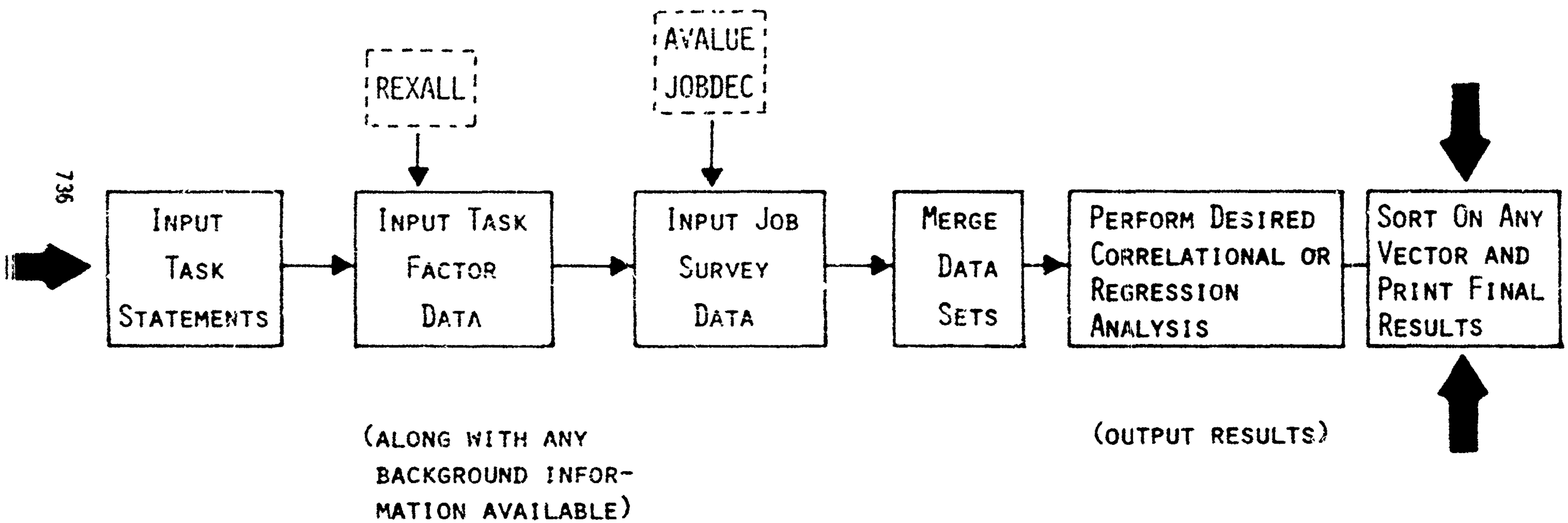
METHOD

The final output form of the report presented in Table 1 was produced in a single job run. All necessary data for the report were input via the card reader, although it would have been possible to access tape or disc devices. The following is a short, highly simplified description of the means by which SAS duplicated the functions of the task-factor programs.

As outlined in Figure 1 the data stream initially consists of task statements, task-factor data, and job survey data. The task-factor data consists of mean task values in punch card format produced by program REXALL. Job survey data is in the form of punched output from program AVALUE and JOBDEC. Normally, JOBDEC produces no punched output other than a "JDC" card in which to reference the Job Description File (to access data such as percent performing). For our purposes, a modification of JOBDEC was introduced allowing punched output of any values created by the program. Background data on raters could have also been input had there been a desire to later reference data based on certain specified rater history characteristics. As the data is input, SAS automatically creates temporary files to be referenced for use later in the program. The files of interest are merged (by task statement) creating the master data set upon which various statistical routines can be performed.

At this point, SAS has performed in a manner analogous to the following CODAP system programs: TSKFAC, FACSTD, FACSPC. Specific regression analyses can be performed on the merged data using SAS's sophisticated General Linear Models procedure along with the request that predicted values be output to the master data set (essentially the role of TSKCOR and PRLFAC). Prior to being analyzed, data may be modified by SAS (squared, cubed, etc.) simply through program statements at any point in the data stream (COMGEN's function). Last, the final report may be printed in any format or sort order desired (FACSUM's function). TSKCAT's function could easily be duplicated either through program generated "1's" and "0's" for specific tasks

Figure 1. Hpothetical Schematic of SAS Input Stream



Data May Be Manipulated, Transformed, or Deleted At Any Time Along The Input Stream.

were no definitions of criticality to restrict the dimensions along which raters' judgments were to be made. Therefore, we decided to try a rating technique that would:

1. Enable raters to compare tasks to one another rather than to a numerical scale.

2. Simplify the judgment process.

3. Provide an operational definition of criticality.

The paired-comparison technique, although used to scale a variety of kinds of stimuli from things to people, has not been used, to our knowledge, in rating job tasks. In the paired-comparison approach, raters are presented two stimuli and asked to judge which stimulus is greater with respect to some characteristic such as size, brightness, beauty.

We have tried the paired-comparison technique in two studies (Boldovici, et al., 1976; Boldovici, et al, 1977). In the first project, two-hundred forty tank gunnery tasks were ranked in terms of criticality, which was determined by the use of the paired-compaison technique. The Tank Commanders serving as respondents were presented with many pairs of target/range combinations. (An example of a pair of target/range combinations is tank at 2000 to 2500 meters, and light-armored vehicle at 500 to 1000 meters.) The respondents were instructed to assume that they had encountered each pair of target/range combination on the battlefield, and that they could not engage the targets simultaneously. They were then asked to indicate which one of the two target/range combinations that comprised each item they would engage first. A criticality score was computed by counting the number of times each combination was chosen as more threatening ("would be engaged first") and dividing by the number of times it could have been chosen (Guilford, 1954). Inter-rater reliability was in the high nineties. Since the rated items varied only in target type and range, the judgments about target threat or criticality were easy to make. The high degree of rater agreement probably also reflected certain learning experiences that the subjects had in common: Tank Commanders receive formal training in assessing target threat. The high inter-rater reliability, therefore, may simply have indicated that all of the subjects had learned "the same things." The second project provided for answering the question whether similarly high inter-rater reliability could be achieved using the paired-comparison technique with a less homogenous sample of armor tasks, where the dimensions for making the criticality judgments were less obvious than target or range, and where the respondents had not received formal instruction in making judgments of the kind required for the ratings.

TABLE 1

COMPARISON OF ESTIMATED TRAINING PRIORITY FOR STATE DIRECTORS AND FIRST-LINE SUPERVISORS

TASK NO.	TASK	ST_EST	PD_EST	DIFF	ST_RANK	PD_RANK	RANKDIF
0009	ISSUE RECEIPTS OF MONEY TAKEN ON CASH BONDS, COPIES OF REPORTS, ETC.	4.82	5.16	-0.33	6	113	-107
0012	PREPARE BAIL BONDS	4.97	5.10	-0.13	17	106	-89
0008	ISSUE DEPARTMENTAL EQUIPMENT	4.82	4.94	-0.11	5	74	-69
F003	ATTEND CITY COUNCIL, CITY PLANNING COMMISSION AND OTHER RELATED MEETINGS	5.06	5.02	0.04	27	90	-63
0015	RECEIVE AND TRANSMIT MESSAGES ON TELETYPE	5.17	5.63	-0.35	88	144	-56

A PAIRED-COMPARISON APPROACH FOR ESTIMATING TASK CRITICALITY

October 1977

Military Testing Association Conference
San Antonio, Texas

James H. Harris, William C. Osborn
and John A. Boldovici

Human Resources Research Organization
Fort Knox, Kentucky 40121

A PAIRED-COMPARISON APPROACH
FOR ESTIMATING TASK CRITICALITY

Training resource limitations demand that choices be made about what to include in training, and what to exclude Agreement seems widespread that training programs should minimally include tasks that are critical to effective job performance and cannot be performed by new trainees. In military training contexts, this reduces to including in training those tasks that are critical to effective performance in combat. Since combat cannot be realistically simulated, a measurement problem immediately arises; namely, how to measure task criticality.

Prescriptive training development literature typically mentions task criticality as an important consideration in determining training content. The literature is, however, vague on the question of how to measure criticality, and silent on the measurement issues associated with criticality estimation.

Conventional training development methods deal with the problem of selecting tasks for inclusion in training in the following way: A job analysis is conducted, resulting in a task list or "inventory." Expert judgment is then used to rate the criticality of each task, usually on some n-point scale ranging from "irrelevant to the job" to "highly critical to mission accomplishment." The tasks receiving the highest ratings are selected for inclusion in training, and those receiving low criticality ratings are excluded or deemphasized. Since the content of training frequently is determined on the basis of criticality ratings, a question arises as to how much confidence can be placed in the ratings. One index of that confidence is inter-rater reliability: to the extent that several raters independently produce similar criticality ratings, confidence in the job-relevance of training content based on the ratings increases. The test-development axiom is directly analogous: reliability is necessary for validity. Applied to training content, the axiom becomes "reliability (of criticality ratings) is necessary for job-relevance (of training content)."

The reliability of criticality ratings that are used for determining training content seldom is reported (McCluskey, et al., 1975; McKnight and Hundt, 1972). In the few instances where reliability has been reported (Ammerman and Pratzner, 1975) rater agreement has been poor—too low in fact for the ratings to be of practical use. We suspected that low-reliability in these studies was due to two important factors. First, tasks were rated on an absolute rather than comparative basis which, among other things, tends to restrict the range of ratings. Second, there

were no definitions of criticality to restrict the dimensions along which raters' judgments were to be made. Therefore, we decided to try a rating technique that would:

1. Enable raters to compare tasks to one another rather than to a numerical scale.

2. Simplify the judgment process.

3. Provide an operational definition of criticality.

The paired-comparison technique, although used to scale a variety of kinds of stimuli from things to people, has not been used, to our knowledge, in rating job tasks. In the paired-comparison approach, raters are presented two stimuli and asked to judge which stimulus is greater with respect to some characteristic such as size, brightness, beauty.

We have tried the paired-comparison technique in two studies (Boldovici, et al., 1976; Boldovici, et al, 1977). In the first project, two-hundred forty tank gunnery tasks were ranked in terms of criticality, which was determined by the use of the paired-compaison technique. The Tank Commanders serving as respondents were presented with many pairs of target/range combinations. (An example of a pair of target/range combinations is tank at 2000 to 2500 meters, and light-armored vehicle at 500 to 1000 meters.) The respondents were instructed to assume that they had encountered each pair of target/range combination on the battlefield, and that they could not engage the targets simultaneously. They were then asked to indicate which one of the two target/range combinations that comprised each item they would engage first. A criticality score was computed by counting the number of times each combination was chosen as more threatening ("would be engaged first") and dividing by the number of times it could have been chosen (Guilford, 1954). Inter-rater reliability was in the high nineties. Since the rated items varied only in target type and range, the judgments about target threat or criticality were easy to make. The high degree of rater agreement probably also reflected certain learning experiences that the subjects had in common: Tank Commanders receive formal training in assessing target threat. The high inter-rater reliability, therefore, may simply have indicated that all of the subjects had learned "the same things." The second project provided for answering the question whether similarly high inter-rater reliability could be achieved using the paired-comparison technique with a less homogenous sample of armor tasks, where the dimensions for making the criticality judgments were less obvious than target or range, and where the respondents had not received formal instruction in making judgments of the kind required for the ratings.

Forty-eight captains, who were enrolled in the Armor Officers' Advanced Course (AOAC) at Fort Knox during the conduct of the project, served as respondents. Twelve forms of a paired-comparison questionnaire were used. The stimuli to be rated in each form were the tasks for one of four crew positions (Driver, Loader, Gunner, Tank Commander) in one of three tanks (M60A1, M48A5, M60A3). The design of each form of the questionnaire can be illustrated by describing how the form for the M60A1 Driver tasks was designed. Seventy M60A1 Driver tasks were identified during the task-description part of the project. The number of possible different pairs of 70 tasks, then, is 70 x 69/2 = 2415. This, of course, would have been too many judgments for each respondent to make. A partial paired-comparison design (McCormick and Bachus, 1952) was used, in which each of the 70 tasks was paired with each of seven other tasks. This partial pairing approach yielded 245 unique pairs of tasks for the M60A1 Driver. The numbers of pairs of tasks for the other 11 forms of the questionnaire ranged from 135 to 280.

The respondents were instructed to assume that they were company commanders choosing crew members to take on a mission in which fire would be exchanged with the enemy. They were then asked to indicate which of two crew members they would choose, based on whether the crew member could do one or the other of a pair of tasks. An example of a pair of tasks for the M60A1 Driver is:

1. Start tank engine.

2. Move vehicle into defilade firing position upon enemy contact.

Criticality values were calculated for each of the twelve sets of tasks by a standard three-step procedure (Guilford, 1954) which placed the twelve sets of values on a similar positive scale. Inter-rater reliability was estimated by correlating scale values for tasks common to the three tanks. The correlations ranged from .55 to .79, with an average of .68. All were statistically significant ($p < .05$).

The paired-comparison technique holds promise as an approach for estimating the relative criticality of tasks. However, the inter-rater reliability estimates and questions about the validity of the results obtained in the two projects raise separate issues for discussion regarding how to generate task criticality estimates that are reliable and valid.

Reliability

The reliability of the criticality estimates obtained in the second paired-comparison study, though statistically significant and probably greater than the reliabilities of criticality ratings in studies using

absolute ratings (Harris, et al., 1975), seems only marginally acceptable, particularly when compared to the results of the first paired-comparison project. The earlier project, however, differed from the later one in several respects which give rise to some tentative operating assumptions on how to generate criticality estimates that are highly reliable. The reliability of the criticality ratings can be expected to increase with:

1. Specificity of the dimensions along which criticality ratings are to be made. To the extent that investigators can create a uniform set among raters as to the dimensions along which judgments are to be made, rater agreement should increase. Without clear specification of the dimensions for making judgments, raters will "make up" their own dimensions. And if these dimensions differ from one rater to the next, rater agreement will suffer.

2. Common learning experiences among raters. The obvious recommendation—that raters should practice making judgments of the kind required by the criticality study--is warranted only when the condition just discussed (specific dimensions) is met. Practice might otherwise simply reinforce idiosyncratic rater behavior and thus reduce rater agreement.

3. The extent to which complete pairings of the tasks to be rated is approximated. The desirability of eliminating the "luck of the draw" in determining which tasks get paired with one another must, however, be traded off against the heavy rater workloads that characterize complete pairings with large numbers of stimulus materials.

4. The number of times each stimulus is rated. Every respondent need not rate every possible pair of tasks, though this may be desirable. Decreasing the workload of each subject can be accomplished in several ways. Partial pairings can be used, with all subjects rating all pairs. Or complete pairings can be used with some of the subjects rating some pairs and not others. Various mixes of the approaches also may be used—partial pairings with some subjects rating some pairs and not others.

The optimal compromises are unfortunately, not known. Examinations would be interesting, of the effects on rater agreement of various reductions (combined and in isolation) in number or proportion of compared pairs, number or proportion of raters rating each pair, and number of observations per stimulus and pair. The generality of the results of such research would, of course, never be fully established. Questions would always remain about the effects of stimulus materials, instructions to raters, rater experience and so forth, on the results obtained. But if confidence is desired in the results of studies that purport to measure the criticality of combat tasks, then additional research on factors affecting rater reliability seems necessary.

Validity

Any study which claims to measure task criticality raises questions associated with the construct, content, and predictive validity of the results obtained. Construct validity is concerned with the extent to which one measured what one intended to measure. Instructions to the respondents should be designed to create a set for judging criticality and criticality alone. But raters' judgments may be influenced by extraneous considerations such as how difficult a task is to learn or perform, or how frequently it is performed on the job. Questions about construct validity will remain as long as reasonable counterinterpretations of the results can be advanced (Cronbach, 1976).

Content validity addresses the extent to which items used in questionnaires represent the universe of items. The issue of how well the universe of subject matter is sampled can never be fully resolved. Resolution would require widespread agreement on the adequacy of the descriptors used to define the universe, and on precise definitions of what constitutes adequate sampling. On the other hand, if a job domain is carefully partitioned into tasks, and all tasks are included in the criticality study, content validity is not a major concern.

Predictive validity is concerned with to what extent would the criticality scores or predictions made from them, correlate with a direct measure of criticality. Establishing the predicitve validity of the results of a criticality study would require correlating the obtained criticality scores with a direct measure of criticality. Obtaining direct measures of task criticality in combat is, of course, out of the question. Intermediate criteria, combat simulations, for example, might

be used in studies of predictive validity. Of course, achieving adequate measurement reliability under simulated combat conditions would be very expensive, though absolutely essential if any important decisions are to be made based on the simulation results.

Concern with the validity of the ratings, though appropriate, may be premature. Reliability issues associated with estimating the criticality of job tasks have only begun to be raised. Given a) that nothing is known about the validity of criticality estimation, and b) choices between results of known and unknown reliability, training developers would seem well advised to use results whose reliability is known. In this respect, it appears that the paired-comparison technique holds promise as a method of rating task criticality.

REFERENCES

Ammerman, H.L. and Pratzner, F.C. Occupational Survey on Auto Mechanics: Task Data from Workers and Supervisors Indicating Job Relevance and Training Criticalness. Columbus, Ohio: Ohio State University, 1975.

Boldovici, J.A., Wheaton, G.R., and Boycan, G.G. Selecting Items for a Tank Gunnery Test (Draft). Fort Knox, Kentucky: Human Resources Research Organization (HumRRO), 1975.

Boldovici, J.A., Harris, J.H., Osborn, W.C., and Heinecke, C.L. Criticality and Cluster Analyses of Tasks for the M48A5, M60A1, and M60A3 Tanks. Fort Knox, Kentucky: Human Resources Research Organization (HumRRO), 1977.

Cronbach, L.J. Test validation. In R.L. Thorndike, (Ed.) Educational Measurement (Second Edition), Washington, D.C.: American Council on Education, 1976.

Guilford, J.P. Psychometric Methods. New York, New York: McGraw-Hill, 1954.

Harris, J.H., Campbell, R.C., Osborn, W.C., and Boldovici, J.A. Development of a Model Job Performance Test for a Combat Occupational Specialty. Volume I. Test Development. Fort Knox, Kentucky: Human Resources Research Organization (HumRRO), 1976.

McCluskey, M.R., Jacobs, T.O., and Cleary, F.K. Systems Engineering of Training for Eight Combat Arms MOSs, Alexandria, Virginia: Human Resources Research Organization (HumRRO), 1975.

McCormick, E.J. and Bachus, J.A. Paired comparison ratings. I. The effect on ratings of reductions in the number of pairs. Journal of Applied Psychology, April, 1952.

McKnight, J.A. and Hundt, A.G. Driver Education Task Analyses: The Development of Instructional Objectives. Alexandria, Virginia: Human Resources Research Organization (HumRRO), 1972.

THE HIERARCHICAL CLUSTERING OF VARIABLES

A PRAGMATIC VIEW

CAPT BARRY P. MCFARLAND ASD/ENECH

ABSTRACT

The purpose of this study was to evaluate the use of cluster analysis techniques to define task clusters as a tool in the Air Force's Occupational Analysis Program. Two Air Force job analysis surveys were used for the task clustering and the invariance of the test clusters were compared between the two inventories. The clusters identified were generally independent and homogeneous. The clusters also showed in a high degree of reliability when the two separate inventories were compared. A number of implications and potential applications for this type of analysis and results are discussed.

I. INTRODUCTION

The purpose of this study was to evaluate task (variable) clusters as a technique and tool that might be applicable to a wide range of Air Force personnel and management problems on which raw data from the Air Force Occupational Analysis Program already exists. The approach was to evaluate task clusters in terms of common characteristics of tasks and to replicate the clustering and evaluate the invariance of the clusters identified to a second independent set of task clusters.

CLUSTER ANALYSIS:

Cluster analysis reduced to its simplest definition is the grouping of like objects into identifiable groups. As such, cluster analysis has conceptual roots that are as old as man himself. Since assessment of similarities and difference among entities is a universal problem, cluster analysis techniques have been developed and utilized in virtually all areas of human thought and scientific endeavor. The modern scientific roots of cluster analysis can be traced to Linnaeus who first developed a formal taxonomy or genotyping based on the grouping of similar characteristics. The mathematical roots in the measurement of similarity is much more recent with origins based upon the work of Pearson (1901) and Spearman (1904). It was not, however, until the advent of the high speed computer that mathematical cluster analysis became a tool of general applicability to scientific study.

The primary impact of cluster analysis has been in what Tryon et al (1970) call "o-analysis," or the clustering of objects. The objects clustered may be individual job descriptions, cells or genotypes. In biology this type of analysis has come to be called numerical taxonomy (Sokal & Sneath, 1963; Jardine & Sibson, 1971). By factor analysts

this type of analysis has been called inverse factor analysis or in the more simplified version the "Q-technique" (Burt, 1937; Cattel, 1952: Stephenson, 1963). As such it has been relegated to a minor role in factor analysis and has even been completely omitted by some authors in the area, (e.g., Harmon, 1967). Thus being omitted, many psychologists, while having a working familiarity with factor analysis, are completely naive of any modern cluster analysis techniques. In addition to psychologists' and biologists' uses of cluster analysis, other disciplines including electrical and mechanical engineering have found utility in object analysis techniques for problems in pattern recognition.

In contrast to object clustering, very little attention has been given to variable clustering. Variable clustering has long been the exclusive domain of factor analysis,except for the data reported by Tryon & Bailey (1970). There are a number of reasons why more attention needs to be given cluster analysis of variables. First, it is not always the goal of the analysis to explain the common variance in terms of the minimum number of constructs. The cluster analysis, particularly the hierarchical clustering techniques, permits the researcher the flexibility of identifying any number of clusters based on a combination of statistical homogeneity of the clusters and empirical meaningfulness. In contrast, factor analysis has as its most serious weakness the need for rotation of axes. Since the goal of factor rotations is to identify factors in terms of "apparently" common psychological constructs it is interesting to note that the most popular factor rotation technique, Varimax (Kaiser 1958) is sadly lacking in its ability to provide sound psychological constructs that have stability and consistency. Although Varimax does provide an optimal solution (i.e., maximum common variance with minimum number of factors), the solutions are frequently of little utility

in providing the researcher with a better understanding of the data. This is as much a criticism of our current psychological constructs as a criticism of Varimax and factor analysis in general. Cluster analysis however does not have the aforementioned analysis problems.

AIR FORCE OCCUPATIONAL ANALYSIS PROGRAM:

The Air Force has used cluster analysis techniques as a part of the occupational analysis program since 1958. The program uses job inventories for the collection of quantitative data obtained directly from job incumbents who describe their job within an Air Force specialty area. In completing the job inventory, each job incumbent supplies identification and background data on himself and then checks an extensive listing of job tasks that are inclusive of all tasks performed within the specialty. In addition to identifying each task performed, the incumbent rates all tasks he performs on a 7-point scale indicating the relative amount of time spent on each task compared to all other tasks performed. The ratings range from 1 (very much below average) to 7 (very much above average) with 4 being a mid-point (about average).

The techniques for developing the job inventory and occupational analysis procedures are reported in a series of research reports dating back to 1958. For the best summary of general procedures for the construction of job inventories, see Morsh and Archer (1967). The past research and continuing experience with job inventory data indicates that these clustering techniques produce highly reliable information about existing Air Force jobs.

The technology that currently exists in the analysis of job inventory responses allows comparison of job types obtained through the cluster analysis of individuals (objects). The tasks or variables are not routinely clustered into independent groupings. Since the total number of tasks for

any given specialty may be extremely large, it would be advantageous to have objective and empirically defined task groupings. Such groups would provide the occupational analyst with task summary information that would benefit the analysis of a specialty code. Also such information could be used to identify possible redundant items in the inventory. As a data bank of tasks is developed, the relative independence and redundancy of items would be available for future inventory development.

Perhaps most important is the potential of identifying task groupings with the same or similar rubrics across different specialty areas. Such groups would allow direct performance comparisons across Air Force specialty codes with a number of common performance measures. The possible applications of these measures are numerous, ranging from the identification of common training requirements, to grade and skill level authorizations, to officer/airmen manning requirements. Each of these is of critical operational, as well as research, interest to the Air Force.

II. METHODOLOGY

The nurse job inventory contained 648 tasks and was originally completed by a total of 2664 Air Force nurses. Because of some computer programming limitations, a sample of 927 cases were randomly selected to complete the task clustering sample. All 648 tasks were used in clustering although only 575 of these were common to Medical Service Job Inventory.

The medical service job inventory contained a total of 600 tasks and was originally completed by a total of 2716 Air Force Medical Service Corpsmen. A sample of 927 cases were randomly selected from the total sample to be included in the task clustering operation.

In the normal occupational analysis procedure, job typing is performed by clustering individuals (objects) into homogeneous groups. The input matrix appears as:

$$\begin{array}{cc} & \text{TASKS} \\ \text{OBJECTS} & \begin{bmatrix} \quad & \quad & \quad \\ & & \\ & & \end{bmatrix} \end{array}$$

Clustering is accomplished by combining individuals based on the similarity of the percent of time spent on tasks.

To complete the task clustering or variable clustering this matrix was rotated so that the following input was used:

$$\begin{array}{cc} & \text{OBJECTS} \\ \text{TASKS} & \begin{bmatrix} \quad & \quad & \quad \\ & & \\ & & \end{bmatrix} \end{array}$$

The clustering procedure used was the same as that reported by Christal and Ward (1967). Thus matrices of size 648 X 927 and 600 X 927 were used respectively for the Nurse and Medical Service input data.

The hierarchical clustering method used combines the two most common items into a single group and proceeds in an iterative fashion to combine one group at a time until only one group remains. Thus, for the nursing tasks, there were 647 iterations before reaching a group size of 1, and 599 steps for the medical service tasks. For the analysis a percent commonality was used. This is a simple transformation of absolute difference. The formula for the procedure used is given as:

$$100 \times \frac{2 \min (X_{ij})\, X_{ij}}{N \text{ (OBJECTS)}} = \text{Percent Commonality}$$

Note, other measures could be easily used in the clustering routine. If, for example, Euclidean distance is used, similarities between the clustering technique and multidimensional scaling become readily apparent. For this initial effort, the simplest measure was employed since no parametric statistics were to be employed from the similarity measures and since previous research has shown that the clusters derived are empirically sound (e.g., McFarland, 1974).

III. RESULTS AND DISCUSSION

Figures 1 and 2 are a summary of the major clusters identified in the hierarchical grouping. The titles for each of the clusters were assigned based on the task statements and previous experience during job analysis on nurses and medical service personnel. The clusters were defined as task families. Each task family consists of a set of tasks which are more homogeneous within that task family than with any grouping outside the task family. Thus, if an individual performed any one of the tasks within a task family, there is a higher probability that he will perform another task in that family than he would perform a task in another specific task family. The tasks comprising each task family are independent and mutually exclusive of all other task families.

Each task in each family was taken and applied to the job analysis data obtained in the nurse and medical service job analysis. Cumulative time spent values were computed for each task family for all task families identified (Nurse and Medical Service). Intercorrelation matrices of the time spent data were computed for sample of 3,115 nurse and medical service corpsmen to determine the commonality between the task families. The intercorrelation matrices are shown in Tables 1, 2 and 3.

Table 1 shows the intercorrelation among the nurse task families based on the percent time-spent in each task family for nurse and medical service corpsman. Note that the highest correlation is .6747 between task families 100 and 351. As shown in Figure 1, these two task families came together at group stage 90. Other than this one relatively high correlation, the rest of the matrix clearly demonstrates a high degree of independence for each of the task families identified. The relatively high correlations found with task family 100 are somewhat inflated because of the large number of tasks contained in this group. Table 2 shows similar results for task families identified from the medical service job inventory. Two high correlations were identified. These were between Groups 349 and 211, and between Groups 94 and 211. Groups 349 and 211, as shown in Figure 2, came together to form Group 89. Thus, the high correlation between these two task families was in fact expected based on the hierarchical clustering of variables. The two task families, 211 and 94, are highly related in that personnel typically performing one set of tasks also perform the other. It can be argued that these two task families should have grouped together in the cluster analysis. The reason they did not group together was that the two groups did not exist at the same stage in the hiercharical clustering. This is considered a major disadvantage of limitation in any acretion method of hierarchical clustering. The advantage that hierarchical clustering provides is that although the tasks in the two groups are independent and mutually exclusive, the similarity between the time-spent values for the two groups was readily identified. Thus, although hierarchical clustering does not provide an "optimal" solution, the occasional resulting discrepancies are easily identified. Aside from the two high correlations already mentioned, the rest of the correlation matrix in Table 2 shows the relative independence of each job type.

Table 3 shows the correlations between the task families identified in the two separate analyses, nurse and medical service corpsman. Unlike the previous two matrices, the tasks comprising any one task family are not mutually exclusive (i.e., common tasks do occur between task families for the nurse and medical service families). This table merely identifies the commonality or uniqueness of each task family previously identified. In Table 3, note that three nurse task families and three of the medical service task families were unique (having less than 25% common variance). From the medical service task families, the following were uniquely defined as belonging to the medical service career field: Group 345 Inventory tasks, Group 319 Admissions and Group 14 Emergency Room tasks. These task families are not only statistically unique from the nurse families but have a great deal of logical and intuitive appeal as well. Nurses do not perform, as a rule, the relatively menial administrative tasks that would be associated with admitting patients to the hospital or performing an inventory of supplies (nor in the Air Force do they have the same responsibilities or near the same numbers of personnel working in emergency rooms). (See McFarland 1975 for comparative analysis of jobs performed)

The task families that were identified as belonging uniquely to the nursing career were Group 283 Irrigation tasks, Group 93 Orthopedics tasks and Group 43 Anesthesia tasks. These task families generally represent specialized tasks for which nurses have been trained and the medical service corpsmen have not been trained to perform. For example, the irrigation tasks are generally performed in an operating room and the Air Force has an entirely different career field for operating room technicians that is totally independent of the medical service career field.

Overall however, high correlations were found between task families that were independently identified from nurse and medical service job inventories.

This is a good reliability check on the technique and clearly demonstrates that, aside from the six unique task families identified, all other task families were consistent across these two similar, albeit separate, Air Force occupations. This stability is a critical element if the clustering of variables is ever to be applied to the myriad of problems for which it shows potential.

POTENTIAL APPLICATION OF VARIABLE CLUSTERING WITHIN THE AIR FORCE

The identification of task variable clusters that are homogeneous within themselves but are independent and mutually exclusive has a great deal of potential benefit in relating task information to many areas of management interest. Potential areas in which task family data might be of benefit include the development of models for the prediction of skill level requirements for airmen positions and the identification of aptitude requirements based on task performance requirements. Closely related to these is the potential for job redesign. This could be useful in identifying task families that require an unusually high aptitude. Two different career fields might be developed, one consisting of a few high aptitude personnel, the other (probably much larger) of lower aptitude personnel.

To test the potential of using task family data to predict skill level requirements for Air Force jobs, a multiple discriminate analysis was computed using job incumbents' self-reported duty skill level (DAFSC) as a criterion against six of the task family time-spent vectors as predictors. The use of DAFSC as a criterion is very limited. It may not reflect the skill requirements of the current job because an individual with a supervisory skill level may in fact be performing a journeyman level job which would not be reflected in his DAFSC. The results of the predicted skill level using standardized distance from the centroid as the measure of best fit are presented in Table 4.

The results, although showing a regression effect toward the mid-point, show the potential of using task family data for prediction of airmen skill levels. Had a better criterion been available, it is felt the results would have been even more striking in showing the potential of using task family data.

CONCLUSIONS:

Many questions are still left unanswered. As mentioned in the introduction, there has been very little psychometric research on cluster analysis. Basic questions still exist as to the optimal measures of similarity that should be used for a cluster analysis, and algorithm provides the most near optimal clustering results.

Cluster analysis is a simple and straightforward analysis technique. It does not require the same rigorous definitions about the data that are required in factor analysis while it still provides the user a certain interaction and freedom in defining the clusters of specific interest. The results of this study clearly show the utility of using a variable clustering technique and the potential that exists for its application to a wide range of research problems.

REFERENCES

1. Abramowitz, M., & Stegun, I.A., 1968: Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, National Bureau of Standards Applied Mathematics Series, Number 55, U.S. Government Printing Office, Washington, D.C., Seventh Printing.

2. Burt, C. L., 1937: Correlations Between Persons. Brit. J. Psychol., 28, 59-96.

3. Cattell, R. B., 1952: Factor Analysis, Harper, New York.

4. Cattell, R. B., & Coulter, M. A. Principles of behavioral taxonomy and the mathematical basis of the taxonome computer program. In M. Lorr & S. B., Lyerly (Eds), Proceedings of the conference on cluster analysis of multivariate data. New Orleans, La: Catholic University of America, June 1967, 5.01-5.38.

5. Christal, R. E., & Ward, J. H., Jr. The MAXOF clustering model. In M. Lorr & S. B. Lyerly (Eds), Proceedings of the conference on cluster analysis of multivariate data. New Orleans, La: Catholic University of America, June 1967, 11.02-11.45.

6. Harman, H., 1960: Modern Factor Analysis. Chicago: University of Chicago Press.

7. Jardine, N., & Sibson, R., 1971: Mathematical Taxonomy. John Wiley & Sons, Ltd., New York.

8. Kaiser, H. F., 1958: The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187-200.

9. Lorr, M., & Lyerly, S. B., 1967: Conference on cluster analysis of multivariate data. New Orleans, La: Catholic University of America, Dec 9, 10, and 11.

10. McFarland, B. P. Job analysis of the Medical Service career field. AFHRL-TR-73-36, AD-775 720. Lackland AFB, Tex: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

11. McFarland, B. P. AFHRL TR-76-52 Comparative Analysis of Nurse and Medical Service Personnel. Lackland AFB, Tex: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, December 1976.

12. Morsh, J. E., & Archer, W. B. Procedural guide for conducting occupational surveys in the United States Air Force. PRL-TR-67-11, AD-664 036. Lackland AFB, Tex: Personnel Research Laboratory, Aerospace Medical Division, September 1967.

13. Pearson, K., 1901: On lines and planes of closest fit to points in space. Phil. Mag., 2, 557-572

14. Spearman, C., 1904: The proof and measurement of association between two things. Amer. J. Psychol., 15, 72-101.

15. Stephenson, W., 1953: The study of behavior: Q-technique and its methodology. Chicago: University of Chicago Press.

16. Sokal, R. R., & Sneath, P.H.A., 1963: Principles of Numerical Taxonomy, W. H. Freeman, San Francisco and London.

17. Tryon, R. C., 1967: Person-clusters on intellectual abilities and MMPI attributes. Mult. Behav. Res., 2:5-34.

18. Tryon, R. C., & Bailey, D. E., 1970: Cluster Analysis, McGraw-Hill Book Company, New York.

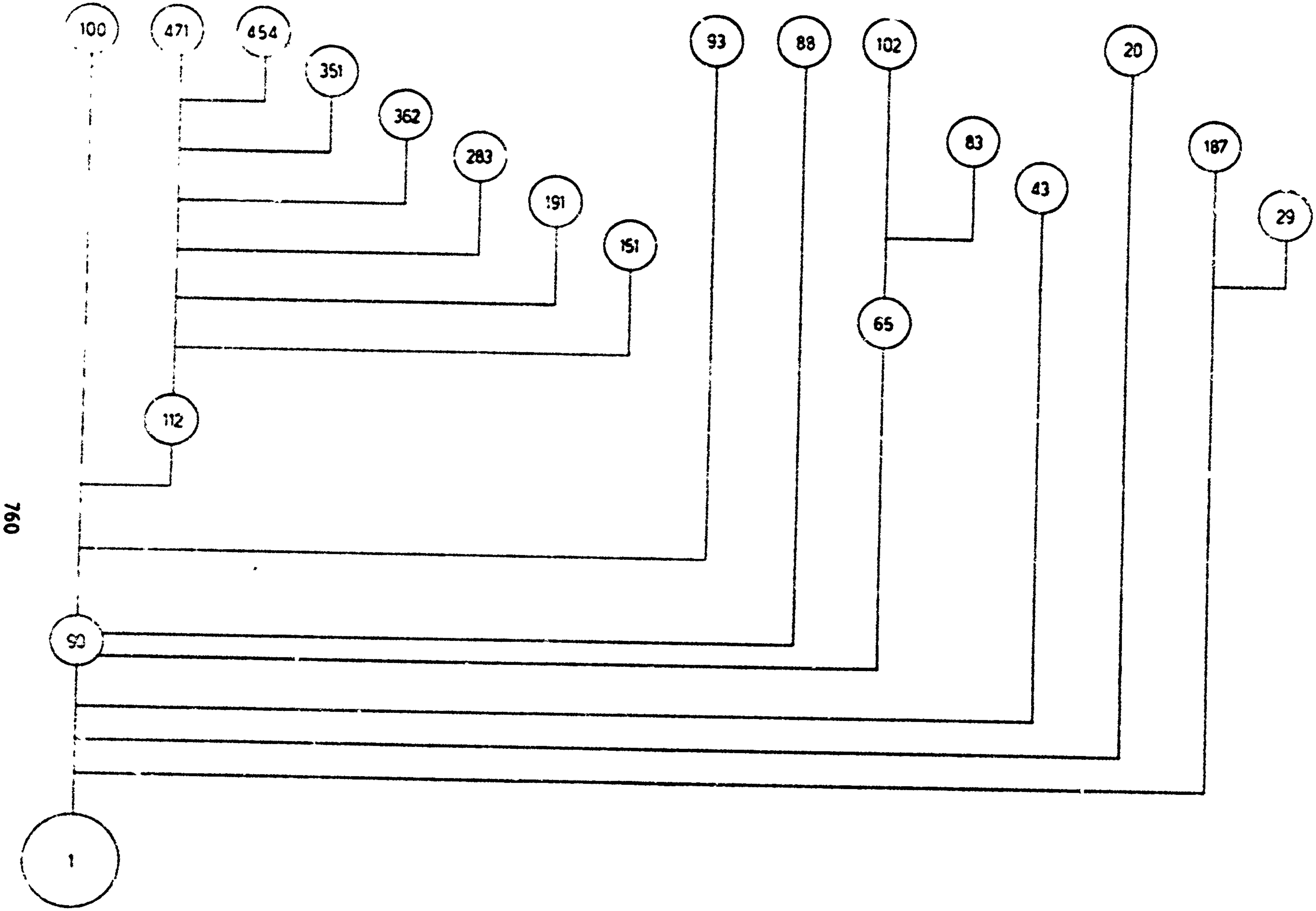


FIGURE 1 - NURSE TASK FAMILIES

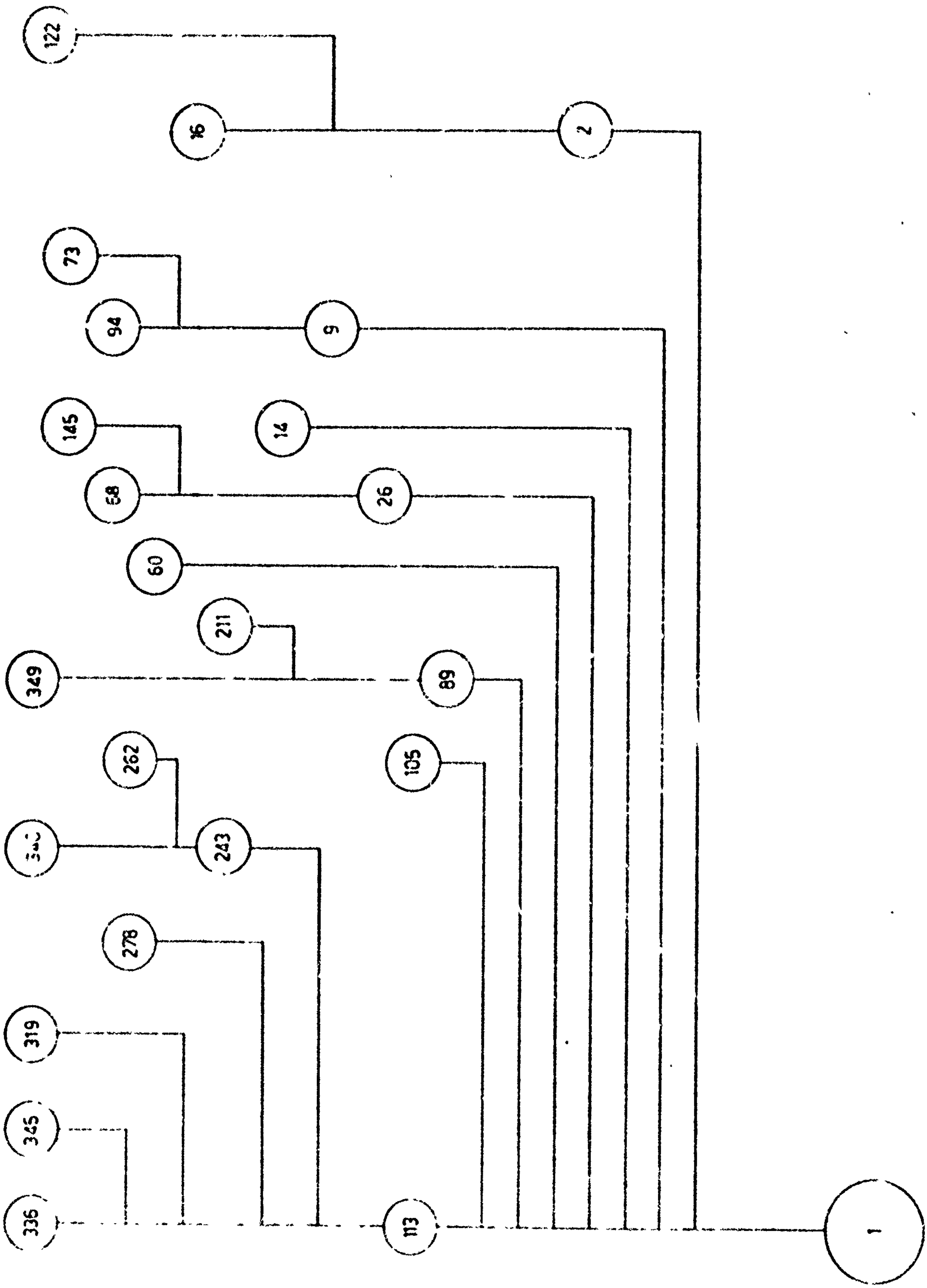

122
2
73
94
9
145
14
58
26
60
211
349
89
262
105
243
278
319
345
336
113
1

Table 1. Nurse: Intercorrelation Matrix of Task Families

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.0000															
2	-.5535	1.0000														
3	-.5695	.4437	1.0000													
4	.6747	-.3082	-.4126	1.0000												
5	-.3636	.1493	.2051	-.3100	1.0000											
6	-.2932	.0255	.1250	-.2916	.1647	1.0000										
7	-.1677	-.0458	-.0469	-.1446	.4893	.2290	1.0000									
8	-.2605	-.1131	.3020	-.2814	.3283	.3206	.3039	1.0000								
9	-.1539	-.1358	-.0421	-.1846	.1252	.4771	.1541	.2629	1.0000							
10	-.2200	.1808	.1406	-.2066	.1262	-.3107	-.1536	-.0981	-.3041	1.0000						
11	.2140	-.3519	-.2780	.0940	-.1685	-.1160	-.0842	-.1083	-.0573	-.0785	1.0000					
12	-.2215	-.0925	-.1136	-.2279	.0602	-.0594	-.0668	-.0811	-.0635	.3119	.3396	1.0000				
13	-.1512	-.0495	-.1079	-.2119	.3221	.1140	.1280	.0061	.1672	.1359	-.0222	.2276	1.0000			
14	.0479	-.0429	-.4751	-.0221	-.2055	.0990	-.0333	-.1645	.1237	-.1287	-.0090	-.0465	.0562	1.0000		
15	-.0485	-.2335	-.1676	-.0713	-.1072	-.0513	-.0492	-.0585	-.0577	-.0877	-.0283	-.0949	-.0314	-.0496	1.0000	
16	.4720	-.4097	-.3794	.1917	-.2357	-.1689	-.1055	-.1685	-.0814	-.1049	.0719	-.1684	-.0747	.0663	.0231	1.0000

Table 2. Medical Service Intercorrelation Matrix of Task Families

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.0000																
2	-.1591	1.0000															
3	.1814	-.0714	1.0000														
4	.3193	-.1424	.0074	1.0000													
5	.2070	-.2392	.1521	.1637	1.0000												
6	.3329	-.1878	-.1133	.1875	.4205	1.0000											
7	-.1222	-.0978	.4279	-.3189	.1555	-.2298	1.0000										
8	-.5504	.1913	-.3199	-.2169	-.3110	-.2744	-.1826	1.0000									
9	-.5186	.1523	-.3324	-.2787	-.2751	-.2212	-.1544	.7210	1.0000								
10	.0055	-.1896	-.0712	.1654	.4736	.4487	-.0607	-.1845	-.1645	1.0000							
11	-.0127	-.0841	.0313	.1346	-.0157	-.0668	-.1895	-.2286	-.1978	-.0035	1.0000						
12	-.0953	-.0863	.1235	.1015	.1246	.0187	.0144	-.1860	-.1411	.0875	.4100	1.0000					
13	-.2657	-.0838	.0971	-.2980	-.0539	-.1789	.5519	-.0555	-.0105	.0193	-.1503	-.0534	1.0000				
14	-.5107	.1031	-.2981	-.2543	-.2605	-.1916	-.1457	.5682	.7240	-.1295	-.1716	-.1147	.0281	1.0000			
15	-.3389	-.0755	-.2721	-.1803	-.1671	-.1223	-.1595	.1499	.1292	-.0969	-.1165	-.0897	-.0534	.1675	1.0000		
16	-.1552	-.0612	-.0458	-.0236	-.0553	.0208	.0574	-.0586	-.0138	.1160	-.0165	-.0056	.2317	.0244	-.0056	1.0000	
17	-.1616	-.0472	-.2029	-.0958	-.0955	-.0377	-.1215	-.0481	-.0142	-.0475	-.0777	-.0695	-.0595	.0130	.0023	.0030	1.0000

Table 3. Correlation Matrix of Nurse by Medical Service Task Families

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-.6442	.5195	.8030	-.4786	.1373	.3432	.1007	.2895	.0712	.0118	-.2950	-.1021	-.0865	-.2974	-.1486	-.4427
2	.1953	-.2261	-.1989	.2993	-.2507	-.1165	-.0759	-.1407	-.0864	-.1542	.1252	-.0101	-.1139	-.0152	-.0402	-.0278
3	-.3877	.4626	.1061	-.2326	.1379	.1549	-.0005	-.0568	.1098	.0796	-.1851	.1862	.1407	-.1230	-.1944	-.3400
4	-.3088	.3716	.7070	-.1620	.1697	-.2232	-.1070	.1199	-.2302	.2664	-.1872	-.0773	-.1193	-.4461	-.1167	-.2491
5	-.3733	.2485	.2299	-.3102	.7326	.3197	.5501	.3433	.1843	-.1747	-.0437	.2038	.2038	-.1938	-.0945	-.2298
6	-.2947	-.0283	.3847	-.2962	.4456	.3318	.3634	.6768	.1196	-.0533	-.1215	-.0960	.0136	-.2418	-.0425	-.1691
7	-.2071	.2471	-.2788	-.1381	.0831	.3637	.0607	-.1183	.4606	-.2075	-.1225	.0332	.2645	.5661	-.1132	-.1861
8	.8697	-.4237	-.4645	.8112	-.3169	-.2707	-.1819	-.2488	-.1509	-.1956	.1411	-.2317	-.1751	-.0240	-.0522	.2852
9	.8350	-.4884	-.4670	.5472	-.2835	-.2124	-.1300	-.2071	-.1218	-.2078	.2295	-.1699	-.0970	.0267	-.0128	.2929
10	-.2040	-.0115	.1066	-.1980	.6172	.0680	.5875	.4497	.0830	.0012	-.1018	-.0260	.1435	-.0825	-.0608	-.1345
11	-.2458	.0281	.0499	-.2360	.1371	-.2284	-.1153	-.1102	-.2331	.7556	.2316	.7270	.1902	-.1734	-.0818	-.1707
12	-.1944	-.0049	.0215	-.1693	-.2148	-.0849	.0043	-.0186	-.0298	.2428	.0723	.6864	.2389	-.1252	-.0749	-.1321
13	-.0121	.0242	-.3705	-.0339	-.0530	.0925	.0470	-.1297	.1774	-.1199	-.0384	-.0501	.1331	.8187	-.0592	-.0346
14	.7262	-.4588	-.4343	.4197	-.2560	-.2021	-.1384	-.1868	-.1151	-.1606	.2199	-.1471	-.0748	.0594	.0036	.3587
15	.3732	-.2955	-.2796	.0862	-.1667	-.1326	-.0850	-.1320	-.0720	-.0388	.0189	-.1271	-.0491	.0427	.0019	.9413
16	-.0089	-.0594	-.1076	-.0720	-.0148	-.0509	-.0248	.0592	.0368	.0585	-.0065	-.0231	.0486	.2175	-.0073	.0092
17	-.0454	-.2306	-.1607	-.0675	-.1056	-.0669	-.0529	-.0549	-.0596	-.0802	-.0769	-.0949	-.0308	-.0542	.9855	.0256

TABLE 4 PREDICTED SKILL LEVEL VERSUS ACTUAL SKILL LEVEL

	ACTUAL		
	Entry Journeyman	Technician	Supervisor
Predicted Entry Journeyman	86.60%	22.73%	2.27%
Technician	12.52%	63.37%	18.18%
Supervisor	.88%	13.90%	79.55%
Total	100.00%	100.00%	100.00%
N=	2141	374	44

Cell Entries = Percentage of Cases

COMMONALITY AND COMPLEXITY AS FACTORS IN SKILLS TRANSFERABILITY

by
T. M. Ansbro, Career Development Group
Naval Education and Training Program Development Center

As armed services' occupational data banks (MODB, CODAP, NOTAP) reach storage levels approaching the monumental, recording of essentially unique and specialized work/duty experience data must eventually taper off; inputs must begin to show massive commonality among elements of job task performance: commonality inherent in the task elements if not immediately apparent in task statements. Irrespective of the historical parochialism of work domains and their attendant descriptive language, close examination of performances underlying tasks draws attention to similarities in logic, application of routine, and manipulation.

Technicians (electronics and others) diagnose, using symptoms-analysis and option-exercise techniques and logic not far removed from such exercise in medicine. Draftsmen draw on paper; carpenters and molders draw on wood; metalsmiths, machinists, and hull technicians draw on metals; others draw on fabrics. For at least part of their work, they all draw: they, and others, also drill, punch holes in, stitch, pin, crimp, nail, screw, rivet, or otherwise attach fasteners to wood, fabric, metal, plastic, paper -- even teeth and bones. Dental tools transcend their designed special application, as witness their use in micro-miniature circuit repair and jewelry- and model-making.

Commonalities discovered in job/task/skill analysis do not necessarily remain at the same levels, nor are they always obviously similar throughout the hierarchies in which they are found. But, when job-common items are identified and their stratification determined, job-unique items then appear in equal clarity.

The importance of commonality and, hence, the finiteness of its distribution, lie mainly in its potential for eventual use in training, manpower management and performance testing. Commonality suggests transferability across and among ratings, NECs, MOSs, and other work categories; it is a factor to be employed in building modularized training programs and tests. Particularly in testing, especially on a broad scale, established and identified commonality in task performance elements makes clear what specific work behaviors are being tested throughout a test or series of tests. Further investigation of the impact of discovered commonalities within and among registered occupational categories may

---

The text of this report is not to be construed as official doctrine of the Department of the Navy, unless so designated by other authorized documents.

confirm the theories of some researchers that there is an inherent symmetry to the distribution of common work items. If so, commonalities abound at the lower skill levels -- job entry or apprentice -- and are most obvious in task statements related to those levels.

With the acknowledgement of skill levels and the recognition that commonality alone has far-reaching implications, hierarchies of tasks, skills, and other associated work data demonstrate the impact of inherent task or skill complexity as well. The inclusion of the modifier "inherent" to describe complexity is for the purpose of differentiating "complexity" from "learning difficulty", two terms which are sometimes argued as being, if not practically interchangeable, then at least reasonably close in meaning, especially within the training community. The distinction in terms becomes important in this regard: whereas the difficulty experienced in learning to perform a task may very well closely match the inherent complexity of that task, learning difficulty can be affected by variances in training methodology, breakthroughs in media application, or by changes in sequencing. Learning difficulty, then, becomes a variable factor. At least relatively speaking, the complexity of a task is fixed, therefore inherent. Operation of a bulldozer, crane, or pipe organ requires simultaneous coordinated use of both hands and both feet; at least in a physical sense, hardly anything can change a complexity factor pinioned by that fact, unless a significant change in the associated machine or manipulation requirement takes place, thereby most likely changing the task itself. Gathering and examining these and numerous other such examples provided mounting evidence that here were numerous opportunities to establish reasonably fixed and almost universally understood criteria (commonality, complexity, and eventually componency and criticality) eventually reducible to quantification and data processing for front-end job/task/skill analysis. This realization led researchers of the Navy Career Training Analysis Group (CTAG)* to a re-examination of the structure of the Navy's world of work, the language used to describe it, and the established taxonomies of job/task/training analysis and the systems approach. That a Gestalt approach was appropriate became evident early: the reason -- the Navy's world of work is a highly complex structure; it does not neatly coincide with the existing categories descriptive of work effort in the other armed services or with the components of the systems approach to training (job, duty, task, task element, etc.). If, for example, "job" aligns with or equals MOS or AFSC, it will not similarly align with a Navy rating. Of increasing importance to the mechanics for conduct of front-end analysis within the Navy, the available components in the Instructional Systems Development or systems approach hierarchies were matched with those divisions of Navy work effort that appeared to be most closely parallel in meaning: job, task, and task element were matched with rating, Naval Enlisted Classification (NEC), billet, and watch for the most compatible pairings. At the outset, job and rating appeared to be a poor match, since the inevitable conclusion drawn from such a pairing would be that there are

*Davis, D.D. & Ansbro, T.M., Occupational Analysis for Navy Instructional Systems Development (ISD): A Matrix Approach. Report to MTA Conference, 1976.

only as many enlisted jobs in the Navy as there are ratings (84), a number clearly inadequate to describe the complexity and variability of scope of Navy work, except in the broadest terms. Within the rating structure are more than one thousand NECs, classifications that tightly define and describe work in specific occupational areas and on specific items of equipment, further aligned with the Navy's principal activity communities: surface, sub-surface, and air. Were job to be matched to rating, the limited compressibility of the available taxonometric hierarchy would most likely force overloading of either the duty or task levels in the resultant inventories. A further complication would be imbalances inherent in realignment of work assignments among the three major activity communities and NEC distribution and range among and within the ratings themselves. For purposes of job/task analysis, it had to be possible to catalog to the level of single discrete work behaviors, or task elements; therefore, among the subordinate work definitions of the Navy, NEC appeared to be the closest match to job, frequently blanketing billet and watch to a satisfactory extent. The compass of NECs is such that there appeared to be some instances of redundant coverage in terms of assumed commonalities among representative items of equipment operated and maintained, some instances of apparent componency and/or personnel input spanning several ratings, some instances of wide variety of technological requirements; but, in the main, NECs appeared supportable and satisfactorily definable as jobs. Coupled with rating descriptions, their numbers appeared to reflect adequately the characteristics, complexity, and variety range of Navy work.

Since rating aligns above job in a vertical matching, it may appear to have been ignored, since job equals NEC, duty equals responsibility grouping of tasks, and task equals task. Rating as a parent assembly of jobs/NECs is affected by the results of job analysis, not necessarily by, or during the course of, the analysis itself; therefore, it is not ignored, it is merely a category level above those taxonometric items used in the analysis, and it is recorded as a parent assembly. Further modification to the taxonometric array was an attempt to maintain a terminology strictly aligned with specifically identified skills behavior. Accordingly, in the concentration on employment of performance characteristics only in conducting the analysis, it proved necessary to drop "knowledge" behavior out of the categories in use. Knowledge behavior was then much more specifically dealt with as identified mental or intellectual skills and mental skill-directed task performance elements; the amended taxonometric categories therefore become more definitive and detailed at the task behavior level and below. "Knowledge", as a familiar catch-all term for sometimes undefined work behavior was minimized, the major bonus being that all the included work information became quantifiable performance data in the computer. The leaner taxonometric array remained in harmony with ISD philosophy, essentially driving testing toward performance measurement

and increasingly orienting training to performance objectives while the door appeared to be opening on consideration of eventual performance testing for advancement and job certification in the fleet.

As indicated generally in the introductory paragraphs, the world of Navy work shows distribution in more than one direction. Complexity, difficulty, componency, and other interrelationships drive the data toward vertical hierarchies commensurate with allocation of responsibilities, tasks, and component skills among job-entry, whole-job performance, upper (advanced) technician levels, and management, herein described in the CTAG effort as: Basic Operator/Technician, Operator/Technician, Advanced Operator/Technician, and Manager. Craft or tradesman (technician, operator, administrator, etc.) boundaries established historically (even sometimes maintained so) and in consequence of weapon, hardware, and platform introduction and change generally blanket the tighter NEC structure which responds not only to hardware and the differing environments of the community (air/sub-surface/surface) overlays, but also to other evidence of uniqueness in work. In such a work matrix, commonality weaves a number of horizontal lines through the patterns. The included technologies tend almost inevitably to group themselves into inherent associations, or families, of work effort, straining at the traditionally or empirically imposed boundaries. Essentially, Navy maintenance is divided into categories (Organizational, Intermediate, Depot) that accommodate manpower management, staffing of maintenance organizations, flow/storage/distribution of spares, and return of equipment to operational status.

The maintenance categories establish clearly <u>where</u> (organizationally and geographically) maintenance takes place. The <u>nature</u> of maintenance <u>performed</u> (or kind of work actually done) falls substantially into four other categories, essentially subsets: preventive, corrective, preservative, and replenishment, only one of which (replenishment) appears to be the exclusive property of a single officially established maintenance category (organizational). In addition, some maintenance personnel (or ratings) perform throughout the entire spectrum of repair without actually fabricating any parts or other items used in rebuilding, reassembly, or repair; others spend much of their working time in fabrication while nominally employed within the maintenance category. The nature of equipment, hardware, materials, and manufacture enters here; for example:

a. Precision - manufactured and tuned components must be replaced in toto on a go-no-go basis, a characteristic of some repair work in electronics.

b. Equipments like engines are reassembled with precision-manufactured parts which may require some honing, fitting, or other alterations during repair.

c. Equipments like parachutes and other items of air delivery/safety may require partial or extensive remanufacture of items because of the nature of their construction and the characteristics (wear, durability, fatigue, etc.) of their materials.

d. Some fabrication of spares may be required in response to emergency situations which occur only on rare occasions and are outside of the normal and predictable maintenance requirements.

e. In other applications, wiring a panel or soldering parts into a circuit, plastering or cementing over damaged masonry, replacing rotted, rusted, or otherwise damaged planks, piping, railing, ladders, gratings, etc. may be considered fabrication; they certainly are considered maintenance.

Maintenance work generally includes operation of support equipment, tools, and instruments, a reasonably probable fabrication of parts or other items (gaskets, seals, shims, etc.), directed modification of the maintained items, some administrative recording, ordering, or reporting, operational testing or run-in of equipment maintained, many inspections and decisions concerning condition, quality, and operational status of items maintained. Therefore, maintenance work or a maintenance designation or assignment appears to include: operation, fabrication, and administration. However, each of these descriptors is a primary work category by itself.

a. A builder primarily fabricates and constructs; he only incidentally maintains.

b. An equipment operator primarily operates designated equipment (earth-moving/materials-handling; he secondarily maintains it).

c. A Personnelman or Yeoman primarily performs administrative functions; he rarely, if ever, maintains anything (although his job description may refer to "maintaining files", it can be considered out-of-context usage).

It is obviously impossible to avoid overlapping and general muddying of the work descriptors. Also, they won't stay put. For instance, medical personnel operate tools and instruments and conduct a great deal of administrative records-keeping. They also maintain their instruments. As indicated in the introduction to this paper, technicians do a great deal of diagnosis, drawing, deciphering, tool and instrument manipulation, and just plain thinking and figuring that may be more alike than they are different.

Ongoing collection of work data continues to demonstrate commonalities across as well as among crafts, trades, or other established divisions of labor, not necessarily at the same levels, nor always obviously similar through the hierarchies in which they are found.

Obviously, in any attempt to catalogue commonality among identified areas, or families, of work activity, the methodology used also points out uniqueness; thereby providing two useful categories. In addition, it is where (at what level of skill) commonalities are found, that forms the major influence on employment of work activity data. If the general groupings of common tasks and skills prove to be essentially symmetrical, that is, if fundamental work actions form the bulk of common items; then, personnel input into a particular work force can be general and broad-based, and the supportive training for appropriate job entry could likewise be general. The most beneficial result of the discovery of such symmetry in the data would be the possibility of grouping work distribution and training factors to provide common apprenticeships or job-entry qualifications in associated fields of work.

The field of inquiry or area from which data input was to be captured was arranged in broad categories of work effort assumed to include significant common characteristics. The principal purpose of such a classification scheme overlay was to find and use associations among the data in such quantities that confident size, and some indications for early payoff from use of commonalities might be realized. Such arrangements of Navy work effort would incorporate presently identified jobs and other classifications into families. The premise is not new. It has been proposed in one form or other quite often. However, as introduced here, it is not a recommendation for a permanent reorganization of the Navy world of work. It was rather intended as facile and appropriate research system machinery designed to effect some initial desired results from the CTAG research assignment and to aid in setting the general parameters for the data base so that it would maintain its general characteristics and projected usefulness in a somewhat uniform manner and without major alteration largely throughout the unavoidably long run of the Navy multi-job/task/skill (and training) analysis.

Data assembly could then be represented by families of work effort, such families characterized by clear similarities in job responsibility; representative equipment, tools and instruments used; similarities in methodology; obviously associated fundamentals of technology (electronics, hydraulics, mechanics, etc.); and what technically expert experientially-derived assumptions could be made to set temporary or experimental boundaries for the families.

In general, the Navy world of work breaks down into the following broadest categories: Administration, Fabrication, Maintenance, Operation, and (purely) Military functions. There are obvious overlaps, of course (fabrication and maintenance, for instance); but these categories do identify general fields of Navy work within which are the familial structures (families) introduced above. The familial boundaries

the field of inquiry was introduced and given a trial run. It appeared to reinforce the initial familial structure assumptions. An equal opportunity for eventual massive consolidation may not occur again, and gains further downstream in the overall run of the system may well be much more modest.

It should be noted that some of the ratings proposed experimentally for classification within the electronics family (figure 1) appear also in the somewhat more loosely assembled electro-mechanical family (figure 2). Such assembly adds further reinforcement to the premise on which accumulated occupational data would be employed.

The primary work categories mentioned earlier (Maintenance, Fabrication, Administration, Operation, Military) are those recognized at the present stage of research to provide reasonably clear boundaries and sufficiently broad areas to include and generally segregate major classes of Navy work effort according to what are determined to be the primary characteristics of the work. The principal value of this further venture into pre-analysis classification was its use in designing individual matrices for the task analysis portion of the entire analysis system. It is an indication of the inherent validity of such classifications for the overall effort that tasks found in maintenance show differing task element sub-structures than those found in such an area as military watchstanding or administration. Also, operational and administrative types of tasks add duty sub-categories and change reference and other requirements on the worksheets. The nature of reference material changes as well. Military doctrinal materials, regulations, and instructions tend to be duty rather than task oriented. Rather than attempt to design a single analysis matrix (and associated data-entry devices like Job Data Worksheet, etc.) to cover all types of task data input to analysis and computer coding entry, it was decided to construct separate matrices for the identified primary work categories.

In summary, pre-analysis classification overlaid on the Navy world of work serves as a road map for the analysis route through ratings and NECs, appears to enhance cumulative and orderly acquisition and early employment of work data, and facilitates ongoing refinement of the tools of analysis. In no way does such pre-analysis classification freeze the accumulated data into categories that may later prove restrictive, unwieldy, or inappropriate. It merely plots the route and sets the course for the ongoing effort. The data base can be cross-coded to facilitate updated task commonality printouts, almost concurrent with task data accumulation. At the present stage of research, this has most recently been accomplished, although not yet in final form. Should category overlaps prove obstructive or assumed associations wear thin, adjustments can be made without harmful effect on the analysis. Restructure is always possible, especially with the insights forthcoming from accumulated experience with the run of the system itself.

engendered by the term "electronics" successfully cover associated methodology, theory of operation, equipment, tools, instruments, references, materials, and assumed technical background. So too, do those described by: communications, weapons control, propulsion, navigation, and detection -- families and family sub-structures of Navy work. They divide (in terms of functional field) generally into operational and maintenance activity, and these distinctions are clear and not obstructive to familial grouping.

Why, when there are already such established Navy work and personnel classification entities as rating, NEC, billet, watch specialty, and community, is it necessary or advisable to include yet another work classification; and an overlay, at that? Because overlaying a familial-structure grouping gathers together those ratings that can be assumed to encompass the associated methodologies, theories of operation, equipment, etc., that can be termed a Navy industrial family. For example, see figure 1, a listing of ratings assembled into the assumptively established "electronics" family. Ratings in this family appear generally to divide into major areas of employments (navigation, detection, etc.), or family sub-structures, which are further divided into two modes: operator and maintenance. These familial structures suffice to set boundaries around twenty-six ratings that function in one or both modes in the area described by electronics.

Conducting analysis progressively across and among the ratings and NECs within this family before moving the research to another group of ratings has the advantage of (1) proving or disproving the assumed-association basis of establishing the family, and (2) making maximum use of commonalities discovered for the purpose of designing common job-entry manpower input and associated training programs. The data base so constructed tends to be homogeneous, and organizationally compresses the distance between initiation of analysis and roughing-out career management and training program recommendations.

Examination of Figure 2 will show a further extension of familial structuring -- an electro-mechanical grouping. At this point in assembling the associated ratings here, no attempt was made to balance the involvement of the ratings among the modes and sub-structures. The groupings, as here demonstrated, represent a first cut at assembly -- an admittedly assumptive grouping intended to enhance data gathering and programmed retrieval with a strong orientation toward examining such commonalities as surfaced. In an experimental grouping across Navy ratings, it was determined that the twenty-six ratings identified as electronics-associated presented most fruitful prospects for early and substantial payoff in terms of projected or at least viewed possibility of consolidation of job-entry training and non-NEC-specific training beyond the first enlistment, with constant manpower management and enlisted career program development opportunities as well. Further, a decision-making model to influence determination of the structure of

ELECTRONICS FAMILY

1. COMMUNICATION

OPERATOR	MAINTENANCE
AC	AT
CT	CT
OS	ET
RM	IC
TD	RM
	TD

2. DETECTION

OPERATOR	MAINTENANCE
AW	AT
AX	AX
CT	CT
DS	DS
EW	EW
FT	FT
OT	ST
ST	TD
TD	

3. WEAPONS CONTROL

OPERATOR	MAINTENANCE
FT	AQ
GM	ET
MN	FT
MT	GMT
ST	MN
TM	MT
	ST
	TM

4. NAVIGATION

OPERATOR	MAINTENANCE
AC	AE
DS	DS
IC	ET
OS	IC
QM	TD
TD	

5. METEOROLOGY/OCEANOGRAPHY

OPERATOR	MAINTENANCE
AG	AG
AM	OS
OT	ST
ST	

6. DATA STORAGE/TRANSMISSION

OPERATOR	MAINTENANCE
DP	DP
	DS

Figure 1

ELECTRO-MECHANICAL FAMILY

1. PROPULSION

OPERATOR	MAINTENANCE
AB	AB
BT	AD
EN	AS
EO	BR
GS	BT
MM	CM
	EN
	GS
	MM

2. MATERIALS & LOAD HANDLING

OPERATOR	MAINTENANCE
AE	AB
AO	AO
BM	CM
EN	EM
EO	EN
MM	MM

3. ACCESSORY EQUIPMENT

OPERATOR	MAINTENANCE
AB	AB
AD	AD
AM	AE
AO	AM
BM	AO
BT	AS
EM	BM
EN	BR
GS	BT
MM	CM
MR	EM
PR	EN
SM	GS
	IC
	MM
	PR
	SM

4. POWER-GENERATING

OPERATOR	MAINTENANCE
AM	AE
BT	AM
EM	BR
EN	BT
MM	CM
UT	EM
	EN
	IC
	MM
	UT

5. PRECISION-MEASURING

OPERATOR	MAINTENANCE
AD	AD
AE	CM
AM	EM
IM	GS
MM	IM
MR	MM
UT	

6. WEAPONS

OPERATOR	MAINTENANCE
AO	AO
GM	AQ
MN	AT
TM	AX
	CM
	DS
	EM
	ET
	FT
	GM
	MM
	ST
	TD
	TM

Figure 2

Symposium: USAF Occupational Measurement Center Programs

Chairman: Colonel James A. Turner, Jr.
Commander
USAF Occupational Measurement Center
Lackland AFB, Texas 78236

This symposium outlines the ongoing programs of the USAF Occupational Measurement Center - construction of occupational tests in support of the Weighted Airman Promotion System and the USAF Occupational Survey Program. These major missions will be outlined in detail and recent trends and developments will be dicussed. In addition, the interaction of these programs will be examined. In some respects, the occupational survey program can be considered as the initial step in ascertaining Air Force utilization policy for each specialty area. Occupational Survey Reports are used in utilization conferences which result in validation or modification of Special Training Standards, Technical Training, Career Development Courses, On-The-Job Training, and eventually in changes to the Specialty Knowledge Tests. Thus USAFOMC is involved in both ends of the process and is in a unique position to make a substantial contribution to the Air Force personnel subsystem.

Presentations in this symposium include:

Overview: USAF OMC Organization and Missions
Capt C. D. Gorman
Management Applications Section

Trends in Test Development
Capt J. R. Johnson
Test Construction Section

USAF Occupational Survey Program
Mr J. B. Keeth
Airmen Career Areas Analysis Section

Management Applications and Special Projects
Major S. D. Stephenson, Chief, Officer Survey & Management Applications Section

The Interface Between Occupational Surveys and Test Construction
Capt David Vaughan
ATC Technical Applications Center

THE USAF OCCUPATIONAL MEASUREMENT CENTER: AN ORIENTATION

by

Captain Charles D. Gorman
Occupational Survey Branch
USAF Occupational Measurement Center
Lackland AFB TX

The purpose of this paper is to provide an overview of the activities of the United States Air Force Occupational Measurement Center, located at Lackland Air Force Base in San Antonio, Texas. The center's mission as depicted by the organizational emblem, is to measure the job and measure the people performing the job. More specifically, the center is charged with conducting Air Force occupational surveys and developing personnel tests in support of the Weighted Airman Promotion System.

The manning posture of the center is unique. There are approximately equal numbers of military and civilian people, with 57% of the center's personnel identified as professional. Most of these professionals hold advanced degrees, primarily in the behavioral sciences.

The center is comprised of four branches. The survey branch performs all work associated with the operational Air Force Occupational Survey Program. The support branch provides the many functions of an orderly room and also maintains one of the largest technical and documentary reference libraries on the base. The test development branch is responsible for major test revision and test research. Support of testing programs and minor test revisions are provided by the test services branch.

The mission of the Occupational Survey Branch is three-fold. First, surveys of airman specialties are accomplished on a recurring basis, approximately once each four years. A new system for maintaining up-to-date task lists for job specialties, called the Current Task Inventory Bank, will provide the center with the capability to survey more frequently if necessary. The second mission responsibility is to survey officer utilization fields. These surveys are accomplished on a selective basis depending upon requests from major using commands or Air Staff agencies. Finally, the branch provides occupational survey data relevant to specific management problems throughout the federal government. These special projects are accomplished on an as-requested basis.

The occupational analysis process involves four steps. The first is the development of a comprehensive list of all the tasks that may be performed by an individual in the specialty being surveyed. The second step involves a validation of the job inventory by subject-matter specialists stationed at operational units worldwide. The third step involves administration of the job inventory to job incumbents, usually through the personnel office at the local installation. The final step in the process involves

analysis of results. Data collected with the job inventory are processed by computer through the use of the Consolidated Occupational Data Analysis Programs, or CODAP, developed by the Air Force Human Resources Laboratory, and are then interpreted by an occupational analyst who prepares the report.

Air Force occupational analysis provides several types of information. For example, specific tasks that personnel perform in accomplishing their job and the items of equipment they use or maintain can be identified. Task analysis provides data that allow the identification of the jobs that are performed by graduates of Air Training Command courses. In addition, the data provide information which identifies the progression of work that occurs within each specialty. Finally, task analysis provides data which can be used to assess the relative difficulty of jobs within a specialty, the relative difficulty of tasks, and the critical nature of individual tasks.

The final occupational survey report is sent to a number of staff agencies who utilize the information in a number of ways. One use of the data is to determine if the existing classification structure is appropriate. Officials at the Air Force Military Personnel Center use the data in conjunction with other information to restructure career ladders. The data are utilized by the other mission element of the center to aid in the development of promotion tests. All of the task analysis data are processed and stored at the Air Force Human Resources Laboratory and are available for their use in many different areas of personnel research. The prime use of occupational survey data in the recent past has been in the determination of training requirements and in support of the instructional systems development process.

To summarize the Air Force Occupational Survey Program, it is a program which provides data which help decision-makers determine what to train; when to provide additional training and what that training should consist of; the classification structure which will facilitate mission accomplishment; and finally, how work should be designed.

Let's turn now to the other half of the center's dual mission, developing promotion tests. The idea of using a weighted formula for promotion purposes was implemented only in 1970. Its purpose was to provide a visible system so the airman could see his relative standing in promotion competition and ensure more equitable promotion opportunities among enlisted personnel throughout the Air Force. Support of the Weighted Airman Promotion System is the job of two of the center's branches.

One key promotion instrument is the specialty knowledge test, or SKT, which occupies the efforts of the center 52 weeks a year. At present, tests for approximately 230 career fields and their shreds are revised at least once annually. We also write the promotion fitness examination, or

PFE, which is administered to 400,000 airmen each year.

The factors and weights that were originally developed to support the Weighted Airman Promotion System concept have recently been revalidated at the Air Force Human Resources Laboratory. These factors have been found to result in highly similar decisions to those rendered by actual promotion boards. As can be seen in Table 1, testing, in the form of the first two factors, accounts for 42 percent of an airman's total possible promotion points. Notice that the first two factors are the only ones over which an individual has control. With this in mind, the importance the Air Force places in having the best possible tests should be obvious.

Non-commissioned officers Air Force-wide have almost unanimously supported the Weighted Airman Promotion System. A major reason for this approval has undoubtedly been the concept of tests "written by airmen for airmen." The center's policy is to solicit only the most highly qualified individuals to write these tests.

Two types of test construction projects are necessary to allow the flexibility to meet Air Force goals of a new test for each promotion cycle. A major revision is completed as necessary, or at least once every two years. The minor revision is completed when a major revision would be inappropriate or not feasible. The center asks for and normally receives particularly good support from the Air Training Command both in the form of career development courses, or CDCs, needed to write the SKTs and the CDC writer or a highly qualified subject-matter specialist from the school. The CDC writer is always an invaluable asset, since the CDCs are used predominantly as the sole source reference for SKT development.

TABLE 1

WEIGHTED AIRMAN PROMOTION SYSTEM

FACTOR	MAXIMUM POINTS	PERCENTAGE VALUE
SKT Score	100	21%
PFE Score	100	21%
Time in service Score	40	9%
Time in grade Score	60	13%
Airman Performance Rating Score	135	30%
Decorations	25	6%
TOTAL	460	100%

Another selection instrument that does not fall into the promotion category but which affects thousands of enlisted personnel each year is the Apprentice Knowledge Test, or AKT. This device is not intended to measure performance on the job. The results are used with other criteria such as experience, training, and supervisory and command recommendations, to determine whether or not the individual has the knowledge necessary to bypass basic technical training in his specialty. The AKT has proven to be quite a versatile instrument over the years. Upgrading to the apprentice level has continued to be the most common purpose for which the test is administered.

Work on a test development project actually begins months before subject-matter specialists arrive at the center. The scheduling section works with the major command to obtain the best qualified specialists for the test-writing team. Team members must be master sergeant selectees or higher in grade.

The cycle begins with team members taking the current tests. This is done for two reasons: First, to allow team members to see what type of questions comprise an SKT; and second, to spot any faulty or obsolete questions. Usually a team consists of four people. Teams representing specialties with shreds may have more team members.

The next step of the test development process is development of a test outline. The outline is a reflection of the major paragraphs in the specialty training standard. The occupational survey is also used at this point to determine how many questions should be written on each major task area.

Each test construction project is supervised and directed by a test construction psychologist who is either an officer or a civilian professional. A major source for questions is the current test; however, a minimum number of new questions must be written. Teams begin by writing a question. They submit it to the test psychologist who either accepts it as is, rejects it, or modifies it. The test psychologist bases his decisions on psychometric principles.

Once the team and the test psychologist are satisfied with a question, they pass it on to the review psychologist, who has had considerable prior experience as a test psychologist. The review psychologist will look at the question and either accept it or send it back to the team for modifications. A review psychologist can look at a question more objectively since he has not been involved in the initial writing process. After he has approved the question, he sends it to be typed. Questions are typed on a magnetic tape selectric typewriter. This procedure provides easy item retrieval and also allows corrections to be made easily. Portions of the test may also go to the illustration department, which is frequently called upon to produce technical drawings for some of the test questions.

After all the questions have been typed, the team picks the best 100 questions out of a pool of 125 that they have developed. The extra 25 questions are called alternates. The 100 questions are placed in the order in which they will appear on the final test and sent to typing where a test manuscript is produced. This manuscript is reviewed during a process called the master review. The subject-matter specialists, test psychologist, and review psychologist are all present during the master review. Each question and its choices are read aloud. It may be necessary at this time to substitute an alternate for one of the original questions. Once the team is satisfied with the test, they sign the manuscript.

The test psychologist and review psychologist again review the manuscript, this time checking for any grammatical or spelling errors that may have been missed. The manuscript is then passed on to the third level of control, a senior review psychologist. The center has eight of these senior reviewers, and they are responsible for monitoring career fields with respect to test development. Once a senior reviewer is satisfied with the manuscript, he passes it on to typing where a camera-ready copy is made. This copy is reviewed for typing errors and then sent to the publishers. After the test has been published, it must be reviewed a final time for printing errors before it is released to the field.

The total project concept is emphasized by the compatibility between testing and training. These two factors are put on a common basis by the use of the same documents as the input source for developments in both areas. As you can see in Figure 1, the outline used for test projects is developed with familiar documents, including the Air Force Manual 39-1 job description, specialty training standard, and occupational survey report. The sole reference a non-commissioned officer is normally asked to study is the most up-to-date version of his career development course. Thus, when he takes his SKT, he should have a feeling of continuity and significant understanding of the material being tested -- much of it being common to his previous technical school training, skill upgrading exam, and subsequent study. Therefore, a degree of compatibility between testing and training has been realized.

The dual missions of the Air Force Occupational Measurement Center also complement one another. Data from the Occupational Survey Program help insure that airmen receive the best and most relevant training for their specialty; and promotion tests developed at the center help insure that those airmen who best assimilate the knowledge of their specialty are the airmen who are promoted. At the Occupational Measurement Center, we like to think of it as "Management through measurement -- the basis for the best."

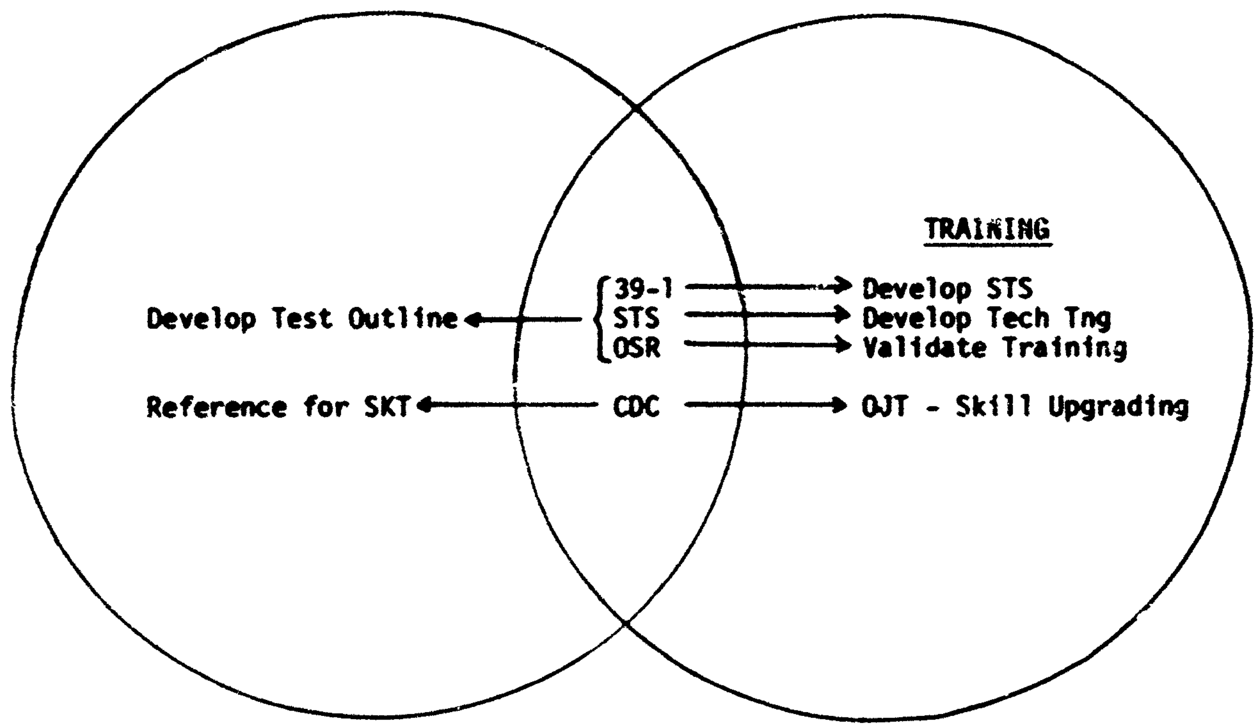


FIGURE 1. Training - Testing Compaibility

TRENDS IN TEST DEVELOPMENT
by

J. Roger Johnson, Captain, USAF
USAF Occupational Measurement Center
Lackland Air Force Base, Texas

I. INTRODUCTION

The test development function of the Occupational Measurement Center is tasked with the development of examinations which are used for occupational placement or promotion within the Air Force. These tests are written by teams of subject-matter specialists (SMSs) selected for their experience in the career field for which they are writing. These teams work in consultation with experts in general test writing technique.

In retrospect over the last two to three years, four trends in this test development function have emerged: first (1), increased attention to conformance of the test construction process with the Uniform Guidelines on Employment Selection Procedures established by the Equal Employment Opportunity Council (EEOC); second (2), increased interaction with the Air Force training function; third (3), an expansion of the test development role to beyond that of the active duty military service member; and finally (4) an increased cost-effectiveness in test development as a response to an austere budgetary environment.

II. CONFORMANCE TO EEOC GUIDELINES

Employment systems across the United States are finding themselves increasingly involved with grievences regarding the validity of examinations used for employment selection and promotion. While Title VII of the Civil Rights Act of 1964, as amended, does not presently pertain to Air Force military personnel per se, we have, nonetheless, taken several proactive research and procedural initiatives to verify and insure that the Air Force test construction process remains fair, unbiased, and in conformance with EEOC Guidelines.

Incorporation of Occupational Survey Data. A study was initiated by which selected SMSs were assigned to relate occupational survey tasks to respective test outlines. Results of the study are expected to facilitate the systematic conversion of job analysis data into test construction. At last year's annual convention, Captain David Vaughan reported a methodology that can eventually be used to translate occupational survey data directly into test outline weights. These weights would be used to determine the number of test items to be written on various specialty knowledge topics within a particular career field. This incorporation of occupational survey data into the test development process will enhance the job relatedness (i.e. content validity) of our examinations.

Attention to Potential Discriminatory Factors. A literature review has been completed regarding the factors that contribute to biased examinations. Subsequently, a study was initiated to review test data for identification of possible ethnic group or sex differences in test scores. Further, test construction teams are instructed to avoid the use of masculine and feminine pronouns that tend to arbitarily assign stereotyped roles. This policy supports the Air Force directive to eliminate sex distinctions in Air Force publications. In addition, a selection procedure has been implemented for members of the teams tasked with the development of our professional military and supervisory upgrade and promotion examinations. The procedure is based upon major command nominations which are screened by the Air Force Military Personnel Center. This procedure ensures an equitable career field, ethnic group, and gender representation among those team members.

Investigation into Spanish Translations. A project was staffed to consider the possibility of translating our tests into Spanish. In researching background information on the subject, many positive and negative aspects associated with translation of tests were identified. Primarily, the concept of testing in the Spanish language for promotion to E-5 and E-7 was considered contrary to the basic philosophy of the promotion system. However, the study did offer alternative solutions related to the problem of language proficiency standards of Spanish-speaking origin Air Force personnel. One such alternative was to identify individuals having trouble in the mastery of the English language very early in their careers and to provide them with the required remedial training.

Promotion Opportunities of Detached Air Force Personnel. At the request of the Defense Communications Agency (DCA), a study was conducted regarding the promotion opportunity of Air Force personnel assigned to duty with the DCA. The concern was that DCA personnel may be obtaining lower test scores than non-DCA airmen. However, the results of the study indicated that there was no significant difference between the groups. Such studies ensure equitable promotion opportunities for Air Force personnel wherever they may be assigned.

III. INCREASED INTERACTION WITH TRAINING FUNCTIONS

The test development function of the Occupational Measurement Center has traditionally supported the personnel classification function of the Air Force, initially with skill upgrade examinations and later with promotion tests. However, in recent years, there has been increasing interaction with the Air Force's training functions.

Specialty Training Standards. Since 1971, test outlines which are developed to guide test writing have been required to be in conformance with the published standards which guide training for each Air Force Specialty. In addition, critiques regarding the adequacy of the training

standards are made by the SMSs who are assembled for test writing. These critiques are forwarded directly to training OPRs for reaction.

Career Development Courses. Whenever available, test reference material is drawn from the Career Development Courses (CDCs) published by the Air Force's Extension Course Institute (ECI). Existing policy supports the development and use of CDCs as the sole source for test references for most career fields. Here too, critiques regarding the adequacy of CDCs are made by the SMSs and forwarded to OPRs for reaction. At a recent conference convened by ECI, the impact of modular extension course materials was discussed. Such materials would have common volumes across many career fields. If implemented, this would enable the development of test item pools for those career fields with common reference materials. ECI personnel have also visited the Occupational Measurement Center as part of a study to determine the extent of correlation between the content and scores of our promotion tests and their end-of-course examinations. Initial findings indicate little content overlap.

Export of Test Writing Expertise. During the last four months, the Occupational Measurement Center has dispatched test psychologists to four technical training centers to participate in local seminars on test development. Rules for item writing, use of item statistics, outline development, quality control, and test validity were topics of discussion. This exchange of expertise is intended to help enhance the overall test development process of the Air Force.

IV. THE EXPANDING TEST DEVELOPMENT ROLE

Although the Occupational Measurement Center's mission has traditionally been mandated toward Air Force active duty military service members, that role is expanding.

Tests for the Guard and Reserve. During the early mid-seventies, the center assisted in the development of a weighted promotion screening system for the Air National Guard and the Air Force Reserve. Similar to the Air Force, this proposed system included a component derived from an individual's score on a specialty knowledge test. Implementation plans and procedures have been developed for adoption of our active duty promotion tests to Guard and Reserve use. Presently the project is in a deferred status due to Guard and Reserve bugetary limitations. Per the last communication received, a re-evaluation of the system is called for during the 1978 fiscal year (FY) budget cycle with intended implementation in FY 79.

Tests for Air Force Civilian Personnel. Since October 1976, the Center has participated in a series of conferences and discussions on the possible use of our promotion tests for civilian wage grade inservice placement actions. Such considerations as test security, the potentially serious impact of test compromise, and the incompatibility between military and civilian occupational tasks led to the conclusion that wholesale adoption

of our military promotion tests for civilian use would be inappropriate. However, the possibility of the center assisting in the development of separate civilian tests appears to be a viable alternative. At this time, dialogue continues on the subject with possible experimental test developments within the next year.

Coast Guard Workshops. Over the past three years, the U.S. Coast Guard has periodically invited center representives to conduct workshops at their training center in Petaluma, California. The workshops have provided training in test writing procedures to Coast Guard instructors in grades E-5 to E-7.

V. INCREASED COST-EFFECTIVENESS

In conformance with the austere budgetary environment in which military agencies are required to function, several actions have been taken to ensure the most efficient and cost-effective test development system possible.

Test Construction Procedural Innovations. In the last year, several procedural innovations in test construction have contributed to increased efficiency. For example, many career fields have common tasks dealing with maintenance management. Given this commonality, a maintenance management item pool has been established to facilitate test writing in that area for those specialties concerned. Other innovations include changing the stagger in the scheduled arrival of teams in order to even the workload on our word processing section, reducing the number of required alternate test items, and excluding from periodic revision those 3-level upgrade examinations which have a low frequency of use. During the last year, the net effect of those innovations has reduced by 5% the duration of test construction projects.

Validation of Test Question Formats. A study was conducted to examine the effectiveness of test question format procedures (e.g. preferences of positive items over negative, multifactor over single factor, situation based over non-situation based, and open stem over closed stem). Experimental tests have been devised and administered and data has been collected. The pending results will ascertain the validity of these format preferences and justification of the time and effort expended in developing them.

Lateral Trainee Testing. Presently, most airmen who transfer from one career field to another attend a basic level technical school if available in the new career field. However, it has been hypotehesized that many of these students may already possess enough basic knowledge and skill to warrant by-passing the basic school. It was therefore suggested that the Apprentice Knowledge Test (given to selected airmen entering the service) be administered as a screening device. Center personnel studied the issue and results indicated that in most specialties, many lateral trainees should, indeed, pass their particular Apprentice

Knowledge Test if mandatory administration was required. While action is pending further evaluation of the criteria for minimum passing scores, the study appears to support the expectation for a substantial savings in formal training resources.

Evaluation of Civilian Contract Test Writing. The center reviewed the possibility of using civilian contract test construction for some special occupational areas. However, the preference for the policy of test construction "by airmen, for airmen" via the subject-matter specialists and the fact that the Air Force cost for test development was less than that proposed by the contractor, contributed to the decision not to adopt the civilian contract proposal.

VI. CLOSURE

The purpose of this paper has been to familiarize the reader with the contemporary issues surrounding test development at the USAF Occupational Measurement Center and to establish the emergence of four trends in that development function, those trends being (1) increased attention to EEOC Guidelines, (2) increased interaction with the Air Force training function, (3) an expansion of the test development role, and (4) increased cost-effectiveness in response to our austere budgetary environment.

THE USAF OCCUPATIONAL SURVEY PROGRAM

by

James B. Keeth
Military Occupational Analyst
USAF Occupational Measurement Center
Air Training Command
Lackland AFB TX

Good Afternoon. I would like to present during this part of our symposium a brief overview of the operational occupational survey program that is conducted by the Occupational Measurement Center for the United States Air Force. The USAF Occupational Survey Program actually began back in 1956 under the guidance of Dr. Raymond Christal of the Air Force Human Resources Laboratory. During the first 11 years of the program, the primary emphasis was placed on development of the methodology required to conduct occupational surveys, including the computer programs necessary for analysis. This research was conducted by Dr. Christal and the Human Resources Laboratory. In 1967, an operational program was set up at Lackland AFB under the auspices of Air Training Command. In the early years of the program, 15 professional, technical, and clerical personnel conducted surveys on 12 enlisted career ladders a year. Today, the program has grown substantially, with 59 assigned personnel and a capability of conducting surveys on 51 career ladders annually, as well as conducting officer surveys, special projects, and electronic principles inventories. Over the last 10 years, we have surveyed over 250 occupational fields involving over 500,000 personnel.

The occuaptional analysis program, as conducted by the Occupational Survey Branch, involves four primary steps. The first step is the development of a job inventory. This consists of a listing of those tasks which may be performed in a particular occupational field. The second step involves validating this task list with subject matter specialists working in operational units. Once the task list has been validated, it is administered to the field. The fourth and final step consists of analyzing the data and reporting the results to the various using agencies.

In developing the initial list of tasks for the job inventory, inventory developers at the Center research all pertinent career field documents and publications, such as the AFR 39-1 Specialty Descriptions, Specialty Training Standards (STSs), and Career Development Courses (CDCs). In addition, previous tasks lists are reviewed for usable tasks. From this research, a tentative list of tasks is developed. During this early stage of the development process, the developer will consult with classification personnel at the Military Personnel Center at Randolph AFB regarding any potential changes planned for the career ladder or other problems that may be of interest. In some cases, functional managers at the Air Staff or MAJCOM level may be involved.

Many times, the occupational survey report will be the first step in a restructuring proposal for an occupational field. An example of this involved the recent occupational survey of the computer systems occupational field. The Air Staff functional manager had received numerous letters from the MAJCOMs regarding weaknesses in the present structure and possible poor utilization of some personnel in the field. To help in coming up with some positive answers to these comments, an occupational survey was requested. Upon completion of the survey, it was indeed determined that personnel were being utilized improperly in some instances and that some restructuring was required. And based on the results of that survey, restructuring has been approved and will become effective next April.

Once the tentative task list is completed, the inventory developers go out to the field to talk first hand with subject matter specialists. This involves visiting the technical training center where training in the specialty is provided. It also involves a visit to several operational units where subject matter specialists are actually doing the job. Once it becomes evident that a fairly complete task list has been obtained from the interviews, the task list is reproduced and mailed out to a small sample of subject matter specialists in operational units in the CONUS and overseas for validation purposes. Once the comments and changes from this field validation are evaluated and pertinent changes, additions, or deletions are made, the task list is printed in a final form and published as a United States Air Force Job Inventory.

Basically, all job inventories consist of two sections. The first section is a set of questions relating to background characteristics of job incumbents as well as to work information. Here we ask for such information as name, grade, command, how long they have been in the occupational field, how long they have been in their present job, and in the Air Force. They are also asked about their job satisfaction and reenlistment intentions. In addition, we obtain information about their work environment, the equipment they use or maintain, the types of aircraft or weapons systems worked on, courses completed, and work section. The amount of information obtainable is almost unlimited.

The second section of a job inventory is simply a detailed listing of tasks which may be performed. When completing this section, job incumbents check those tasks which they perform in their present job. Once they complete this procedure, they go back and rate each task according to the amount of time they spend performing it relative to all other tasks being performed.

Once the job inventory is published, the first two steps in our survey program are completed. Step three involves the administration of the inventory to career field members. Local consolidated base personnel offices receive the booklets and administer them to personnel specified by us. Names of job incumbents who are to receive the booklet are obtained from the Uniform Airman Record file provided by the Military Personnel Center. Our sample is a random stratified sample by skill level, command, and job locations. In occupational fields with 3,000 or less incumbents, we survey the total population of job incumbents who have been on the job at least six weeks or longer. Where there is 3,000 or more job incumbents, some percentage of the total number is obtained. Normally this will range up to 30 percent depending upon the diversity of the field being surveyed.

Job inventories are returned to to the Center where they are scanned and reviewed for completeness and accuracy. It is very important that incorrectly filled out booklets are identified and returned to the job incumbent for reaccomplishment. Also, our survey analysts review the returned booklets to make sure that we are receiving returns from all skill levels, commands, units, and locations. This careful sampling and checking of the returned booklets insure that the final sample will adequately represent the total occupaitonal field population.

Once the returns are reviewed and accepted by the survey analyst, step four of our analysis process begins. The booklets are scanned on an optical scanner and the background information is keypunched, and the data is input into the Human Resources Laboratory's UNIVAC 1108 computer. Using the Comprehensive Occupational Data Analysis Programs, or CODAP as it is commonly known, the data are analyzed and a final occupational survey report is written. As many of you already know, CODAP is simply a series of highly complex computer programs that are used to reduce the large amount of data obtained into a more manageable form.

In analyzing the data, the central focus by the survey analyst is to look at how this occupational field is structured. One unique strength of CODAP is its capability of grouping job incumbents only on the basis of task similarity and producing what is called a cluster merger diagram. From this diagram, the analyst can determine the job structure as it actually exists out in the field. This powerful tool allows us to use this information as a foundation from which we can look at the career field documents, such as the AFR 39-1 specialty descriptions and the specialty training standard, and evaluate if they are accurately and realistically describing what job incumbents are actually doing in the field. These documents are often times based on what is believed to be done in the field and how they are believed to be structured. Our data, through this grouping process, provides information as to how jobs are actually performed and how they are organized. Often times this preconceived idea of jobs does not reflect the actual work performed (Driskill, 1975).

In addition to looking at job structure, the analysis phase can provide other useful information. It can yield a comprehensive listing of tasks performed by personnel in the field. This listing can be broken down into any given group performing each of the tasks. It can show how may people perform each task, as well as the percentage of time spent by those members who perform the tasks. In addition, task and job difficulty data is collected for each occupational field surveyed. Interrater agreement is very high for this data, and we only report data when agreement exceeds .90. Very few sets of ratings have fallen below this criterion to the surveys we have conducted over the years (Driskill, 1975).

Once our data has been analyzed and an occupational survey report has been written and released, the data is put to use by the Air Force in a variety of ways. The data is often used by classification personnel to look at career field structuring, in which the present structure is validated or restructuring is recommended. The data is also used by our SKT branch to aid in developing promotion tests. One of the primary issues today involves having valid criteria on which to base many of the tests administered. Our data can help provide that criteria. Dr. Christal and his staff at the Human Resources Laboratory also use our data for personnel research. Occupational data is also used in the instructional system development (ISD) program to analyze system requirements, to define education and training requirements, and to conduct evaluation of instruction.

But perhaps the most important use today of our data is in determing training requirements. In todays environment where the training dollar is tight, it is all too important that training be geared only to what the person will need to do his job effectively. In this regard, the emphasis today is placed on determining how job incumbents will be utilized in first job assignment, identify those tasks for which the probability of performance by airmen in this first assignment is high, and provide initial training on these tasks. Therefore, our data is useful in designing initial courses that train only for the first job as well as providing valuable information for what to include in follow on training.

This then is a very brief look at the Air Force occupational survey program. We feel that the program has great potential in helping solve many of the problems faced today by the Air Force. More and more people are turning to us for help and we are taking steps to provide this assistance. To help provide us with the capability to take on new challenges, Dr. Christal and his staff are continually providing us with new and improved techniques. All in all, ours is a growing program with a bright future. Thank you for your time.

REFERENCES

Driskill, Walter E. "Occupational Analysis in the United States Air Force," paper presented to the instructional system Design Conference, Keesler AFB, Miss. Oct 1973.

Driskill, Walter E. "Occuaptional Analysis in the United States Air Force," paper presented at the Task Inventory Exchange National Symposium on Task Analysis/Inventories, Ohio State University, Columbus Ohio, on November 18, 1975.

MANAGEMENT APPLICATIONS AND SPECIAL PROJECTS

Stanley D. Stephenson
Rodger D. Ballentine

As the USAF Occupational Survey Program and the awareness of it grew, many special one-time only requests from a variety of agencies started being received. These were handled within the existing structure simply by diverting resources from the normal mission to accomplish the special request. A case in point was the Electronic Principles Inventory (EPI) whose development was presented in a paper at the 1976 MTA and elsewhere.

In addition to special requests officer surveys had not, for several reasons, been a dominant factor in the overall occupational survey program. By 1974, however, the normal program had grown to the point where we were ready to start surveying officer career areas on a more permanent basis.

By 1975 it had become evident that the special projects and officer surveys were an integral part of the survey program and that manning should reflect this. The Officer Survey Management Applications Section was created and staffed in 1976 to handle projects that were above and beyond the normal enlisted career field occupational survey program. The section was staffed primarily with experienced analysts and developers; in order to achieve full manning, however, less experienced personnel had to be added later. Among the less experienced personnel added to the section were an educational specialist and a social psychologist, both of whom increased the versatility of the section to examine unique problems. In fact it has become obvious that such diversity of background adds dramatically to the flexibility of the section.

The first special project undertaken by this new section was to define the job performed by weapons officers flying tactical aircraft. Tactical Air Command training managers requested data to develop training programs for these officers.

Several surveys were developed to gather information about tasks performed, task difficulty and criticality, and job knowledges requiring training. Traditional interview development procedures were primarily used but a unique method was used to validate the job survey task list. Rather than mail the survey instrument to job incumbents for validation, experienced weapons officers came together with survey developers in a working conference to review and validate the task list.

The job survey was administered to all job incumbents and additional task factor surveys were administered to smaller groups of experienced weapons officers. A separate survey was also administered to assess the requirement for formal training of knowledges unique to the fighter weapons area. Survey returns were processed and analyzed using the Comprehensive Occupational Data Analysis Programs (CODAP) package. These data provided valuable insight into the jobs performed and training required for weapons officers. Overall, the success of this project proved that occupational survey techniques can be adopted to meet the objectives of a requesting agency.

Of all the experiences we gained from our first major special project, two stand out and have become a part of our "look out for" list. The first, and the more critical for application, is how the user plans to make use

of occupational data. At the conclusion of the FWIC project it became clear that the requesting command simply was not fully aware of how much information can be gleaned from occupational survey data nor were they fully cognizant of the usefulness of much of the data. Consequently we had to spend time educating the user on our project. To prevent this situation from recurring, we now actively educate the requesting agency starting with the first meeting. This takes the form of briefings on how the data can be used (often in ways unknown to the user); inclusion of user representatives during the project life, especially during the analysis stage; and an agreement on what user office will handle the data and who will be the continuity link from start to finish.

The second major learning experience involved the use of operational requestor personnel in the final validation of the job inventory. This greatly added to our confidence in the job inventory and, in part, helps overcome the uniqueness of each special project. It also proved advantageous in one other way. By having experienced operational personnel involved in the final decisions about the job inventory, it increased the sellability of the inventory to the survey respondents in the field who more often than not had never heard of the occupational survey program. However they normally know of the experienced personnel who helped in the final validation, and by citing their names we greatly added to the quality and quantity of survey returns.

Besides the FWIC we are currently engaged in a wide variety of other projects. The EPI, mentioned earlier, is now winding down after receiving heavy emphasis. For instance, in April 1977 we mailed over 10,700 EPI booklets to all those career ladders receiving SP training who had not yet

been surveyed. Quite naturally this massive effort severely taxed our capability and the capability of the field. Nonetheless we did turn the project around and now have delivered EPI reports on approximately 90% of the career fields involved.

Several other projects warrant mentioning because of either their scope or uniqueness. We are presently involved in an analysis of three reporting identifiers for the Air Force Systems Command. Reporting identifiers are a system of classifying airmen who perform widely divergent tasks; our involvement will truly test our ability to describe job structure. The Defense Intelligence Agency has requested we analyze the job performance of their school graduates. This project is important to us not only because it is a DOD agency but also because it will deal primarily with executive type tasks. Such will also be the case with our Professional Military Education project for Air University. This project, being conducted for the Leadership Management Development Center, will analyze the management, leadership, and communications behaviors performed by both NCOs and officers. Finally our job analysis task list development project for the Federal Procurement Institute is important to us because of its scope (over 20,000 federal procurement positions) and because it is a federal agency versus an Air Force or DOD agency.

Our increased involvement with executive type tasks coincides with the growth of our officer survey program which has served to highlight the importance of being able to measure executive behaviors. We have just completed an analysis of the Security Police career area and are in the

administration stage for two other officer surveys, one for Weapon Controllers and one for Space Systems Analysts. We have found that officer surveys are not merely an extension of enlisted surveys; rather they are a process unique of themselves.

It is obvious that the Management Applications and Officer Survey Section has succeeded in its stated purpose. Perhaps it has succeeded too well, for we find that as the awareness of the value of our consulting grows we simply receive more requests for assistance. We view this with encouragement for it is a validation of how we can assist in critical management, classification, and training decisions. As we learn and grow with each project, we only add to that capability.

The Interface Between Occupational Survey and Test Construction

David S. Vaughan, Capt., USAF
ATC Technology Applications Center

The USAF Occupational Measurement Center has two primary missions—conducting occupational surveys and developing Specialty Knowledge Tests (SKTs). These two missions are certainly distinct; in other armed services they are accomplished by different organizations. However, the test construction and occupational surveying activities within the USAF Occupational Measurement Center have contributed a great deal to each other. The purpose here is to discuss some examples of this cross-fertilization between occupational surveying and test construction.

First, some examples will be given of contributions which occupational surveys are making in test construction. The Specialty Knowledge Tests written at the USAF Occupational Measurement Center are designed to measure job knowledge in particular Air Force job specialties. These tests are developed by teams of subject-matter specialists—senior NCOs on temporary duty status—along with the test construction experts of the Center. Occupational surveys provide extremely good information concerning what airmen actually do in each of the various job specialties and therefore can be useful in determining the topics to be covered on Specialty Knowledge Tests. Since the early days of Air Force occupational surveying the survey data has been made available to test construction teams. This data has been useful. However, test construction teams found the occupational survey reports, which contain a great deal of information, difficult to use. Several years ago, special occupational survey computer printouts were made available for test construction. Those printouts contain simple listings of all tasks in a particular job specialty. Also printed are the task difficulties and percent members performing and percent time spent on each task by airmen in the target population for each test to be constructed. Such printouts present the information which is most useful for test construction in a compact, easy-to-understand format. This printout format, which was developed in close consultation between test construction and occupational survey personnel, has facilitated use of survey data in test construction.

Recently, USAF Occupational Measurement Center personnel have been working on even more effective ways of using occupational survey data in test construction. This author presented a paper at last year's Military Testing Association conference (Vaughan, 1976) concerning one of these efforts. The goal of this effort, which is still in progress, is to develop methods of automatically converting occupational survey data into numbers of test items to be written on each topic. Under procedures currently used at the Center, the numbers of test items to be written on each topic are determined by the subject-matter specialists, based on the occupational survey data and on their professional judgments.

The research has used policy-capturing methodology to find an equation for combining occupational data into numbers of test items to be written on groups of tasks. The data gathered so far show that subject-matter specialists' judgments concerning numbers of test items can be predicted quite accurately from survey data. However, the results also show that different equations are probably necessary in different job specialties. Further research is necessary to investigate the usefulness of a simple, perhaps unweighted, equation across specialties and concerning identification of groups of specialties which can use common equations.

USAF Occupational Measurement Center personnel have also been conducting research on another approach for using occupational survey data in test construction. In this approach, occupational data is used to compute a testing priority index for each task. Tasks are eliminated which are performed by insignificant numbers of airmen in the target population of the test to be constructed. Test construction teams are given a list of the remaining tasks in order of testing priority. The subject-matter specialists study this list of tasks and determine which tasks are suitable for inclusion on a paper-and-pencil, multiple choice test. These tasks are covered on the test, and the numbers of items written on each task are based on the testing priorities of the tasks. Several variations of this method have been used with test construction teams and have been successful. However, further research is necessary concerning algorithms for computing testing priority and for removing tasks from the listing to be given to the test construction teams.

Both of the methods for systematic use of occupational survey data in determining test content have merit. In recent months, a synthesis of these two methods has been worked out. However, this synthesis has yet to be tested. Regardless of which method eventually becomes operational, more systematic use of occupational survey data in test construction will result in a significant time savings for the test construction process. Furthermore, test content validity will be more systematically assured.

The procedures just discussed use occupational survey data for content validation. Occupational survey methodology is also being used to gather job performance data. In a project currently being conducted by Center personnel, an occupational survey task list is being used to gather supervisory ratings of job performance First, an airman indicates which tasks he or she performs. Then, the airman's supervisor rates the airman's performance on each task performed. Hopefully, these task performance ratings will be less contaminated with halo effects than overall ratings usually are and therefore will be more useful in test validation. Furthermore, this approach will allow test scores to be related to not only <u>how well</u> tasks are performed but to <u>what</u> tasks are performed as well. Survey returns are currently awaited. Therefore, the success of this procedure cannot yet be determined. However, it seems likely that occupational survey methodology will make a significant contribution to gathering job performance data for use in test construction.

Occupational survey data can be used in validating individual test

items. The first step in a procedure for item validation is to identify all tasks which relate to each item on a test. The amount of experience on tasks related to an item can be compared to the probability of passing the item. This procedure has not yet been implemented at the USAF Occupational Measurement Center. However, preliminary planning is in progress for a test of this procedure, which is another promising application of occupational survey data in test construction.

Several important contributions which the occupational survey program has made in test construction were outlined above. In return, test construction has made important contributions to the occupational survey program. Subject-matter specialists play an important role both in the development of a task inventory and in analysis of survey results. Subject-matter specialists are an important source of tasks for an inventory. Ordinarily, Center inventory development personnel make extensive TDY trips to other bases in order to get input from subject-matter specialists. However, many subject-matter specialists come to the Center for test construction. When scheduling permits, these test construction personnel are used in developing task lists, thereby reducing the need for Center personnel to make expensive trips to other bases.

Subject-matter specialists can also play an important role in the analysis and interpretation of an occupational survey. One step in the analysis of a survey is establishment of the relationship between the task list and important Air Force training documents such as the Specialty Training Standard (STS) and the Plans of Instruction (POIs). Use of subject-matter specialists can be an important part of this step. Again use of subject-matter specialists brought to the Center for test construction duties can result in significant economy and efficiency. Subject-matter specialists can be helpful in interpretation of occupational survey results. The subject-matter specialists on a test construction team, who have a wide variety of backgrounds and experience in a job specialty, can be particularly useful to survey analysts in data interpretation.

While occupational surveying and test development are two distinct missions, the above examples demonstrate that these two activities are complementary. At the USAF Occupational Measurement Center, each has benefited from the other. Because in the Air Force these two activities are accomplished within one organization, the author believes that the cross-fertilization between occupational surveying and job knowledge testing has been greatly facilitated.

Reference

Vaughan, D. S. Prediction of Test Outline Weights from Occupational Survey Data. Proceedings of the 18th Annual Conference of the Military Testing Association, 1976, 435-460.

THE STATE OF THE ART IN JOB TASK ANALYSIS

Dr Robert Pulliam

INTRODUCTION

Over the past two and one-half years, Kinton, Incorporated has been working with AFHRL collecting learning difficulty data on selected tasks in various career fields in the Air Force. These data were collected using field teams of expert observers and task-anchored rating scales. The task-anchored scales were developed and used to provide a "yardstick" for measuring learning difficulty across similar Air Force career fields, using field teams as the mode for collecting the data. In addition to the rating data, other significant products of this effort were procedural guides for using the task-anchored rating scales. The emphasis of this paper will be on the development of the task-anchored scales and on our success with the use of field teams. But first I would like to provide you with some background leading to this effort.

BACKGROUND

The Armed Services Vocational Aptitude Battery (ASVAB) is used to assess the aptitude of new recruits in the Air Force, but there has been no similarly objective measure of job aptitude requirements. Preliminary survey evidence indicates that some job aptitude requirement levels are out of alignment. As a result, the most talented applicants may not be receiving assignments to the most demanding jobs. This in turn could lead to increased training costs, job dissatisfaction and fewer reenlistments. Thus, a requirement existed for research for a means to assess job aptitude requirement levels for tasks, jobs, and job types in the Air Force.

The USAF began this research with task inventories which described the job content of career fields in terms of specific tasks and other quantifiable characteristics. Included were initial estimates of task learning difficulty. These estimates were gathered during an occupational survey, and consist of learning difficulty ratings by NCOs for tasks in their own specialties, using an adjectival difficulty scale. These data are useful as estimates of relative task difficulty within specialties, but cannot be compared across career fields.

This then was the premise for the development of a common scale by which task difficulty both within and across specialty areas can be measured. To this end, AFHRL has sponsored the development of task-anchored scales for electronic and mechanical aptitude requirements and is in the process of developing a task-anchored scale for general and administrative aptitude requirements.

SUMMARY OF TASK-ANCHORED SCALE DEVELOPMENT

The objective of each of these efforts was to develop a 25-level scale to be used to rate tasks in terms of "learning difficulty." Learning difficulty is defined as "the time required to learn to perform a task satisfactorily."

Each scale was developed using a three-phase effort. The first phase required the selection of 600 tasks across fifteen (15) career fields. These tasks were evaluated by a field team of eight observers who assessed each task at work sites. Each member of the team then rank-ordered the 600 tasks in terms of learning difficulty. AFHRL used the ranking data to generate a mean rank ordering of the 600 tasks. This list was divided into 25 intervals, with two tasks selected from each interval to form a task-anchored rating scale. The scale is now representative of both task types and of learning difficulty across the career fields.

The second and third phases of this effort involved the rating of 60 tasks from each of a selected group of career fields. In these phases, the tasks in each career field were assessed at work sites by two field teams of six observers each. Each team member then rated the tasks using the 25-level scale by comparing the task to be rated with tasks listed on the task-anchored scale. These data are ultimately to be used by AFHRL to assess aptitude requirements for entry into the sampled career fields.

PROCEDURAL GUIDES

It became apparent during the assessment and rating process that it would be necessary to develop a procedural guide, in order to ensure consistency between teams in the criteria used to assess learning difficulty, and in the methods for use of the scale.

The guide was developed in two parts. The first part describes the assessment and rating procedures, and is based on guidance from AFHRL and on Kinton's collective experience in managing the assessment teams. It addresses novice panel members who are presumed to be experienced in the USAF world of work, but who have no psychometric training.

The second part provides a detailed explanation of each task, focusing on the skills or knowledge which must be learned to perform it. The guide was developed after detailed study of each task, principally with the aid of subject matter experts (SMEs) at USAF technical schools. Detailed descriptions were prepared for each task on the scale, noting conditions of performance, criterion standards, specific skill or knowledge required, and any circumstances which tended to mitigate or increase the difficulty of learning the task.

The procedural guide was validated by a field test using three teams to rate 60 tasks in each of four career fields (two electronic and two mechanical). One team had prior experience using the task-anchored scales and rated the tasks in each of the four career fields. The other teams consisted of persons who had no prior experience in using the scale, but a broad general knowledge of the USAF world of work and specific competence in mechanical or electronic career fields (supervisor equivalence). One of these teams rated the tasks in the two electronic fields, the other rated the tasks in the two mechanical fields.

The rating data from the novice teams shows an interrater reliability coefficient of from .940 to .960, and a coefficient of from .910 to .950 for the experienced team across the four career fields. The two teams as a group had reliability coefficients of .97 to .94 across the four career fields. We interpret these data as suggesting that the task-anchored scale procedure is a reliable method, and that the Procedural Guides are effective means to replicate the method.

USE OF FIELD TEAMS

One or more field teams were used in each phase of a task-anchored scale development. Each team typically consisted of from six to eight ex-military personnel. The majority of the team members were retired Air Force NCOs with experience in one or more of the career fields being sampled. Additional conditions for selection included: a requirement to have worked as a supervisor of journeymen and to have had five or more years' experience in the military. Thus, the field teams represented a high level of Air Force/military background, plus hands-on and supervisory experience in one or more of the career fields being assessed.

Each team assessed the task sampling as a group through visits to appropriate Air Force bases. Time constraints precluded the physical observation of each task; thus, the team evolved an interview technique with tours of the work facilities for task assessment. The interviews were conducted as a group with two or more personnel from each career field with each member of the team encouraged to ask questions. The interviews were designed to identify those factors which affect the learning difficulty of each task and to determine to what extent these factors play a part in the learning of that task. After completion of the interview and tour of the facilities, the teams would typically reconvene and discuss the information gethered at the interview. These discussions were used to ensure that the boundary conditions of each task are defined, that all of the factors have been accounted for, and that any biases of the personnel interviewed have been taken into account. Lengthy and often heated discussions occurred at these sessions as the biases of the individual panelists also became known. At no time were the actual ratings of the tasks to be discussed at these meetings.

The ratings were then performed independently by each team member the same day as the interview was performed. Data to date, using the scales, indicate a median interrater reliability coefficient of .955 on over 1500 tasks across 35 electronic career fields and .940 coefficient on over 1500 tasks across 38 mechanical career fields.

Conclusions on the Use of Field Teams

Our experience indicates that the advantages of using field teams to collect data of the type described far outweighs the disadvantages. The most significant disadvantages are the costs of a team or teams for travel and the problem of consistently maintaining an experienced team over long periods of time.

The advantages are that the data are collected at the source by actual understanding of the task with observation of the equipment and worksite. The team members not only get a worksite interpretation of the task through the interviews, but also often observe or participate in actual task performance. For example, one team spent a day in a missile silo observing demonstrations of task performance. In addition, various team members have operated devices from flight simulators to an electrical lineman cherry picker. Additional advantages accrue when two teams are fielded in that data are collected from two different sites in each career field.

The success in using field teams on these projects is a high recommendation for the use of teams in future efforts to collect human factors data.

DEVELOPMENT OF A NAVAL ENLISTED PROFESSIONAL DEVELOPMENT INFORMATION SYSTEM (NEPDIS)

David L. Hannaman
Kinton, Incorporated

October 1977

Paper Presented at the 19th Annual Conference of the Military Testing Association

The United States Navy's Chief of Naval Education and Training (CNET) intends, insofar as possible, to deliver a trained man for every billet at minimum cost. A major obstacle to this objective is that required training decision and development information is not available in a manageable form and timely manner. Needed is not only training objective data, but human management data that will support decisions and development processes about who, when, where, what, and how to train--decisions and processes implied in the Interservice Instructional Systems Development (ISD) Model. For this purpose the Navy has embarked on the design of a sophisticated information system, i.e., the Naval Enlisted Professional Development Information System (NEPDIS).

This paper will present a concept upon which the CNET is dependent if full realization of the Naval Education and Training Command's (NAVEDTRACOM's) ISD effort is going to be realized. The Navy in this author's opinion, cannot expect to exploit ISD to its fullest potential unless an information system, such as NEPDIS, is developed.

PROBLEM

Like the other services, the Navy has adopted and is in the process of implementing ISD - a systematic method of designing, developing, implementing and evaluating the total learning process in terms of specific objectives of the learner. ISD represents a major advancement in training technology and all services have the potential of realizing great benefits by its implementation. However, ISD as it presently exists in the services, represents a "what should be done" policy. It does not, in most instances, specify <u>how</u> each phase or step is to be accomplished.

The paramount and fundamental problem ISD presents the Navy is one of data management. The CNET is responsible for the design, development, management, and evaluation of 2,500 enlisted courses of instruction. Following the ISD model, these 2,500 courses involve approximately 7,000,000 tasks, 10,000,000 terminal learning objectives (TLOs), 11,000,000 TLS criterion referenced performance measures (CRPMs), 20,000,000 enabling learning objectives (ELOs) and 22,000,000 ELO CRPMs. Not only does ISD result in the generation of this data, but it necessitates access to this data on a recurring basis. For example, step 4 of phase I of the ISD model dictates that whenever a training requirement is identified, not only all Navy but all DOD training will be investigated to determine if there is an existing course or portion of a course that will satisfy the training requirement.

If one does exist, a tremendous training development savings could be realized and redundant training avoided. But how does one conduct such an investigation? One approach would be to search existing training by learning objectives and/or tasks. Even if the data were accessible, could an adequate search be conducted manually in a cost-effective manner? Probably not.

Most of the effort that has gone into Naval course development in the past ten years cannot be re-identified in any orderly fashion. That is to say that it is not known specifically which tasks are trained in which courses. Additionally, an audit trail (implied in ISD) is lacking, through which tasks might be traced from stated training requirements, through design and development, coverage in a course, and graduate performance in the fleet....then back to analysis for any reason. It is highly likely that training to specific tasks appears many times in many curricula....and this is probably necessary in some instances, but probably an unnecessary redundancy in others. If the Navy is going to manage ISD and to employ what it has to offer in a highly effective manner, it must have an audit trail through each existing curriculum, with each element of that curriculum linked to all others, so that the data can be located for comparison and update.

The responsibilities for CNET to develop and maintain courses continue to mount. Navy studies have recommended cutting back on the numbers of NECs. This would have made training more feasible and more affordable by permitting core (or common) courses, feeding specific or finger courses in support of NECs which have common skill requirements. But instead of decreasing NECs, the Navy has gone from 677 in 1968 to 1179 in January 1976. This number of NECs may be necessary and based upon valid requirements, but tremendous savings might have been realized in course development costs had it been possible to identify the common skill requirements among related NECs and to combine them into few core courses. There is little need here to address diminishing resources; this is a common condition today in civilian organizations, as well as in DOD. What must be done, then, is to manage ever-diminishing resources better, so that it is possible to respond to ever-increasing responsibilities with logic, based upon quantified data.

RESOLUTION

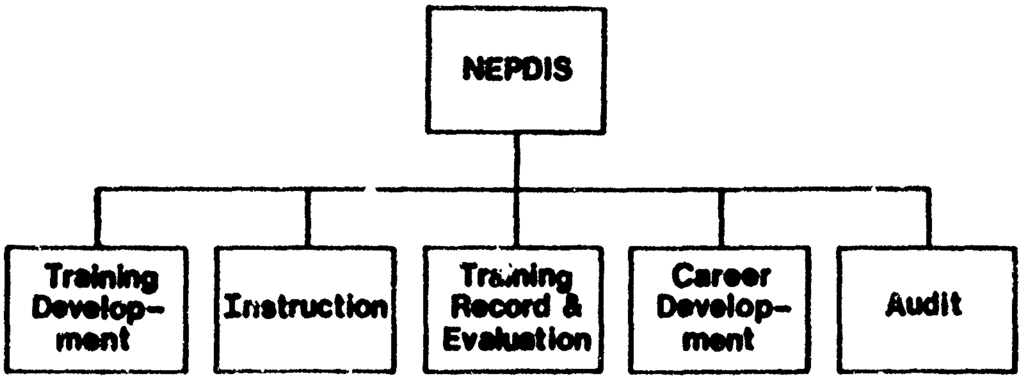
I believe that it is essential that an automated system such as NEPDIS be developed, to provide NAVEDTRACOM with the resources to exploit the opportunities of the ISD model fully and to analyze the multitude of required data, critical to training development and in order to:

- o eliminate unnecessary training;
- o achieve reductions in both training development and actual training time;
- o coordinate the development of training to eliminate duplication of training development, as well as actual training activities;
- o increase the efficiency of training development activities;
- o develop the most cost-effective career training ladders for enlisted personnel;
- o efficiently evaluate work efforts and sequential output material in order to discover and correct deficiencies in a cost-effective manner;
- o ascertain the actual cost associated with the development and conduct of training.

NEPDIS SUBSYSTEMS

To satisfy the objectives for which NEPDIS was conceived, five subsystems or processes are required, as illustrated in Table 1.

TABLE 1



NEPDIS SUBSYSTEMS

The Training Development Subsystem will provide a means to store, manage, and quantify job data in a manner that will enable the identification of commonality, skill levels, and complexities. This subsystem will also maintain a record of all training development activities including their current status, decisions, and the individual or agency responsible.

The Instruction Subsystem will provide the means to record all Naval enlisted instructional programs including non-resident and OBT as well as resident programs, developed and presently being developed. All training material and literature, as well as the criterion referenced performance measurements associated with each instructional program, will be included in these records.

The Training Record and Evaluation Subsystem will support internal and external training evaluation by providing a flexible and readily accessible means for storing and analyzing evaluative data, as well as supporting the selection of study samples and providing a means to access sample populations. It will maintain a record of all Naval training (resident, non-resident, and OBT) acquired by Navy enlisted personnel, as well as a record of skills and knowledge obtained by traditional and non-traditional civilian education, vocational schools, and so forth. Criterion referenced performance testing results for each training or education course completed will be maintained and readily accessible for training evaluation.

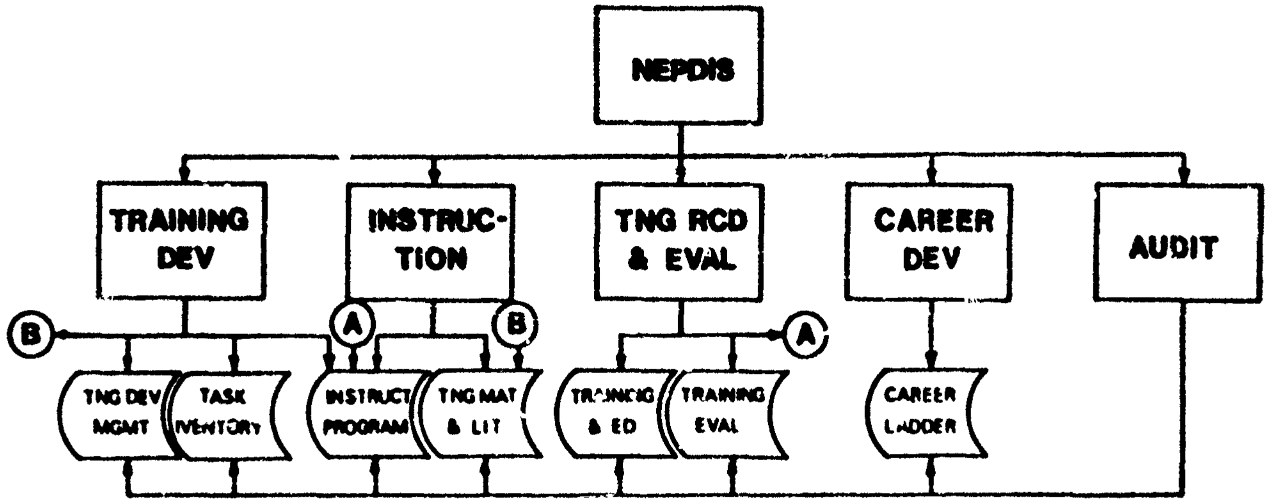
The Career Development Subsystem will provide a means for identifying Naval enlisted career ladders. It will reflect the position of an individual within a given career ladder, and of the career options open to each sailor. In the process, this subsystem will identify, for the Navy, the most cost effective career paths for enlisted personnel.

The Audit Subsystem will provide the means to manage the entire training development system by ascertaining the full impact of changes external to the training community (hardware modifications, operating practice, doctrine, etc.), as well as the internal impact of training decisions. This subsystem will provide an automatic alert to proponents of training activities, when those activities are impacted either by changes in the manpower requirements or in systems hardware.

NEPDIS DATA BASE

The NEPDIS data base described represents a conceptual design only. No doubt several revisions of this data base will occur during the development of NEPDIS. The paradigm below illustrates the NEPDIS data base as it is presently conceived. It should also be understood that, although on-line storage symbols have been used to denote or illustrate these files, it is not the intention at this time to specify the choice between on-line or off-line (batch) storage or processing. These symbols are used merely to identify files - not storage or processing modes.

TABLE 2



NEPDIS
SUBSYSTEMS & DATA BASE

The Task Inventory File will maintain a record of all enlisted task inventories with supportive data that will permit the identification of task commonality, complexity, and skills. "Skills" in this context would include enabling mental and physical skills as well as terminal physical skills.

The Training Development Management File will maintain a record of all training development activities including data such as:

(1) the identification of a task

(2) the decision whether or not to develop training (for a task)

(3) the method of training (resident, non-resident, or OBT)

(4) training design and development

(5) instructional programs

(6) training literature and materials

(7) performance measures

(8) projected and actual training development milestones

(9) the organization and individual responsible for training development activity

(10) the current status of a given training development activity

This file can be viewed as the catalyst for an automated training development PERT.

The Instructional Program File will maintain a record of all Naval enlisted resident, non-resident, and OBT training programs developed or currently being developed. For each program, it will record the associated learning objectives, performance measures, and a course synopsis. This file will provide for an audit trail that can link instructional programs to tasks, training literature and materials, and career ladders. This file will provide support not only to the training development community, but to the schools as well. It can be viewed as an ERIC for the Navy enlisted schools.

The Training Material and Literature File will include a record of all training literature (instructor guides, programmed instruction texts, rate training manuals, etc.) and training materials (films, video tapes, audio cassettes, viewgraphs, slides, etc.) whether developed or presently under development, for Naval enlisted training programs. This file too would be conducive to an audit trail between training literature and material to tasks, learning objectives, performance measures, and so forth. It will support not only the training development community but the instructors as well - an ERIC for school staff.

The Training and Education File will provide a central, comprehensive education and training record for all Naval enlisted personnel. This file will include biographic data, current assignment, high school, college, and vocational education courses completed, as well as degrees or diplomas. It will record USAFI and DANTES courses, and formal Naval training (resident, non-resident, and OBT) to include the results of criterion referenced performance measures. Equivalency examinations and scores will also be included.

A segment of each individual's training record will be reserved for training evaluation data. This portion of the records will be used by the CNET, N5 for depositing training evaluation data for analytic purposes. Once an analysis is completed, this data would be deleted and the file segment could be used once again for the same purpose. This will enable CNET, N8 to have ready access to the entire Naval enlisted population for purposes of training evaluation.

The Training Evaluation File will be a flexible, multipurpose working file, designed so that most training evaluation data can be stored in it and evaluative analyses performed against it. This file should not be confused with the Training and Education File. The Training and Education File will enable training evaluators to access and store information about specific individuals. The Training Evaluation File will be used as a depository of information extracted from the Training and Education File. Evaluative analyses can then be performed on data from the Training Evaluation File in a much more efficient and cost-effective manner than would be possible using the Training and Education File, due to its volume.

The Career Ladder File will record all Naval enlisted career ladders with a capability to identify the grades associated with each career ladder step, the core and finger training required to achieve each of those grades, and information concerning where training may be acquired and when during a given career continuum that training is appropriate.

SUMMARY

The NEPDIS is achievable. It will pay for itself. Furthermore, in the long term it will be unavoidably necessary. Very soon the Navy will not be able to keep up with technology, or to man a competent Naval establishment by using existing ad-hoc methods to decide what will be taught, to whom, and when. Tomorrow's Navy will no more be able to identify its training needs through use of human judgment and hand-processed paperwork than it could manage its logistic system by those same methods. The problems are similar. Every billet in every ship generates a continuing and unique requirement for electronic parts, paint, pipe and paper clips. When those do not arrive on time, the systems go down; if too much is ordered or stockpiled, the costs skyrocket. Those same billets each require unique inventories of skills, knowledge, and experience, yet training developers and designers have never really been able to meet those requirements,

except by guesswork. Large automated systems have been developed to support the logistics community, but so far, no such resources have been applied to the equally complex, important, and costly activities of training.

VALIDATION OF CLASS 'A' SCHOOL TRAINING AGAINST JOB CONTENT

Richard S. Lanterman
U S Coast Guard Headquarters
G-P-1/2/62
Washington, DC 20590

David A. Bownas
Personnel Decisions Research Institute
821 Marquette Avenue
Foshay Tower
Minneapolis, MN 55402

Content validation is a process for showing that some procedure (a test, work sample, training course, etc.) is a representative sample of a relevant domain of activities. In this study, our object is to develop a procedure for determining if the material presented in a Coast Guard A school covers the information required to perform a representative sample of the domain of tasks defining the corresponding job (rating).

The first step in the process is to identify the target domain of tasks. This was relatively simple, since the Coast Guard has recently completeed task analysis surveys for the three ratings we are considering. We will use the task lists emerging from these surveys to represent the task domains for the three ratings, although we will make some checks to ensure that the lists are up-to-date. The information available for each task includes the proportion of personnel performing it, the relative time spent per task, and its average difficulty rating. In addition, we will survey a small sample of personnel from each rating to obtain estimates of each task's criticality (the threat to the overall mission if the task is not correctly performed), and whether each task is performed at a helper or doer level.

Our current plan is to combine the criticality and time spent data to obtain an overall importance index for each task. This will probalbly be done by either adding or multiplying the time spent and criticality ratings, producing an importance index where either high criticality or high time spent can result in a moderate to high overall importance value.

In the next research step, we will identify the class A school training content. We hope to specify approximately fifty (50) curriculum elements or training topics that are relatively

discrete and internally homogeneous for each rating. The school for one rating seems to be broken down rather nicely already, with fifty-three (53) substantive topics covered. The other two are less than ideal, with approximately twenty-five (25) broad weekly topics, and approximately one hundred (100) specific topics at the next level. We are currently meeting with instructors to identify fewer elements at an intermediate level of specificity. In identifying these curriculum elements, we will consider the Coast Guard Enlisted Qualifications Manual, course curriculum outlines, and suggestions from class A school training personnel.

After school content has been specified, the validation of course content will be accomplished by mapping the curriculum against the task to determine whether the most important tasks are receiving the most emphasis in training. This mapping will be accomplished through a series of judgments.

Class A school instructors will make three evaluations. First, they will estimate the extent to which performance on each task is emphasiszed during formal training. Next, they will identify the three or four curriculum elements in which each task is predominately trained. Finally, they will indicate how much emphasis each curriculum element receives during training, primarily as a function of the amount of time spent on each topic or element.

A sample of recent class A school graduates will make an additional set of evaluations to serve as a reliability check on the instructor's data. Since instructors may have a tendency to overestimate what is taught in their courses, we will ask recent graduates to indicate how well they can currently perform each task. We will compare their mean ratings with instructors' ratings of the emphasis given each task to determine whether some tasks may actually be receiving less attention in training than the instructors believe. Where such discrepancies occur, we will request some substantiation of the instructor ratings.

Class A school records will provide a second source of information about curriculum element emphasis. Results of weekly examinations, and final course results are recorded on Student Record Forms for each student. By analyzing these, we can determine empirically how a curriculum element, or a group of curriculum elements for a week, contributes to a final course grade, and hence to whether a student is promoted into a specialty. We will compare these results with instructors' ratings of curriculum element emphasis to determine whether elements are weighted in the final course grade in the same way that they are emphasized in class. Both types of element weight estimates will be considered in evaluating course content

validity.

Training content validity will be assessed from these data in three steps. First, each task will be "assigned" to the curriculum elements where instructors feel it is primarily taught. A task may be apportioned across several curriculum elements according to the extent to which it is taught in each element. The result will be a matrix of curriculum elements by tasks, indicating the extent to which performance on each task is taught in each curriculum element.

In the second step, the matrix entries under each curriculum element will be weighted by the curriculum element's emphasis (either the emphasis it receives in training, or its weight in determining final course scores).

Finally, these weighted matrix values will be summed across all fifty (50) curriculum elements for each task, to obtain one index of how much emphasis each task receives in the training, and one estimate of how much each task is weighted in the final course grade.

These two task training profiles (across the five hundred (500) or so tasks in each rating) will be correlated with task importance data from the task analysis survey, to ascertain whether the task training emphasis profiles match the task importance profile. For ratings where this correlation is high, we will conclude that the tasks emphasized in course content are the same as those most important for successful job performance. When the correlations are low, we will compute variances between training emphasis values and importance values to identify tasks that appear to be over or under-emphasized in training. Lists of such tasks will be provided to curriculum developers, so they can decide whether the course content should be changed to reflect task importance more closely, or whether there is some additional factor which warrants the unusual emphasis given these particular tasks.

Although the need for insuring the validity of training content has been stressed for some time, and although informal procedures exist for developing course curricula from task lists or Qualification Manuals, we believe this is one of the first procedures for quantitatively evaluating the content validity of an existing training curriculum; and we believe the methodology will provide a valuable contribution to the area of personnel training and evaluation.

FINDING A SYSTEMATIC METHOD FOR DERIVING OBJECTIVES FOR FOREIGN LANGUAGE TRAINING

Francis A. Cartier
Defense Language Institute Foreign Language Center

The views of the author of this paper do not purport to reflect the position of the Army or the Department of Defense.

The first principle of instructional systems design requires analysis of the duties and tasks that the graduate of the training will actually have to perform. From the information gathered in that analysis, the objectives and criteria for the training are determined, and all other system development processes, including the tests, are targeted specifically on those objectives and criteria.

In recent years, the Services have performed job and task analyses for a great number of military jobs and acceptable procedures have been worked out for tabulating the analytical information and developing objectives and criteria for the necessary training. As a result, the training has been made more job-relevant and a great deal of training time, previously devoted to irrelevant knowledge and skills, has been eliminated.

The Defense Language Institute has been exploring the applications of instructional systems theory to foreign-language training since about 1971 and has been thoroughly committed to it for about two years.

However, from the very beginning of this effort, certain fundamental problems arose which are yet to be solved.

This paper will not present solutions to these problems, but will describe some of them and discuss our effort to find solutions to them.

Our first difficulty in converting to instructional systems design is the great size and variety of the Defense Foreign Language Program. We teach about 120 different courses at the Foreign Language Center, in about 30 different languages. I say "about" in both cases because the requirements change from year to year. Many of these courses are nearly a year in duration.

The second difficulty arises from the fact that the students in some of our longest courses--the Basic Courses--are not all being trained for the same type of job. DLI Basic Courses do not train personnel for any particular job, but train them to perform part or all of their jobs in a foreign language. As a Defense school, we have students from the Army, Navy, Air Force and Marine Corps, all of which have slightly different ways of identifying the jobs and stating the requirements.

We have, therefore, an enormous number and variety of jobs to analyze. Practical considerations make it necessary for us to trade off some precision in targeting the instruction to the individual job in favor of generalizing the objectives of training for optimum applicability to several student populations.

Fortunately, we do not bear the responsibility for job and task analysis for each and every one of these students. The great majority of our trainees come to us from the security services. After several years of cooperation with the National Security Agency and the cryptological training systems of the Army, Navy and Air Force, DLI is now receiving fairly precise information about the jobs under their control. There are still some unanswered questions in this job field, but these questions are gradually being worked out. The result is that we are getting acceptable terminal skill objectives for the language training of security service personnel and also a considerable amount of information about the enabling objectives. We will not, therefore, discuss this area of language training further today.

But for the students who go to other commands and agencies, we have no systematic method for gathering similar information. This was a major finding of the Army Linguist Personnel Study (commonly referred to as the ALPS) conducted by DCSPer in 1975 and 1976. The ALPS study helped define the problems somewhat, although narrowly. It deals only with the Army, of course. And it deals predominantly with administrative problems rather than with technical problems--as this paper has also done so far.

Now, however, let us turn to the technical problems, since these are the ones that will be of greatest interest to testing specialists in an audience such as this.

The terminal skill objectives for a foreign-language course are typically stated in terms of the various communication behaviors the graduate is required to perform. Such objectives are centered around verbs such as reads, translates, speaks, converses, listens, and so on.

The enabling objectives for an instructional system leading to these kinds of skills are, on the other hand, mostly stated in terms of the specific vocabulary, tenses, case endings, idioms, technical jargon, and so on, that the student must master in order to perform the communication skills.

We are necessarily dealing, then, with two different domains: the domain of communication skills and the domain of linguistics. Defining and circumscribing each of these domains separately presents problems which are beyond the present state of the art. But what is still more difficult is reconciling or combining the two domains in a way that is meaningful for the writer of terminal and enabling objectives and for the designer of the criterion-referenced test that will be used to determine whether the student is sufficiently trained for adequate performance on the job.

If you are teaching Morse code, to take a simple example, the entire domain is comparatively easy to describe. It is quite finite. Each enabling objective can be readily listed and its achievement can be measured with some precision.

Furthermore, the performance test can include virtually every element and significant condition, so the instructional system can be designed quite specifically to prepare the trainee to pass the test. One cannot say the same thing about a foreign language.

Because the desired product of language training is also performance, applied performance testing is highly desirable. This point was made at the 1975 MTA conference in Indianapolis by Dr. Thomas Sachse of the Northwest Regional Education Laboratory. He went on to say, however, that a search for such measures by the Clearinghouse for Applied Performance Testing produced nothing. That is not strictly true today, but shows how new the idea of performance testing is in the field of foreign language training. If it were easily done, there probably would have been at least something to be found in 1975.

The vocabulary of any language is so large that not even a native speaker of the language knows it all. Obviously, we do not expect the trainee to learn more words than a native speaker knows. The question is, "Which words shall be omitted?" It might seem logical to omit those that are not known to the <u>average</u> native speaker. Unfortunately, we have been unable to locate that critically important individual. Every native speaker seems to know a somewhat different sample of the total vocabulary of the language, depending on his or her education, profession, experience, etc. All right, then, let's take the vocabulary that most educated native speakers have in common.

That should be sufficient for our trainee: after all, it seems to be sufficient for the native speakers. Unfortunately, that common vocabulary is also far too large for any practical instructional system.

So let's try a different approach. We know that many people can get along fairly well with a limited vocabulary by choosing their words carefully. If you have mastered about 2,000 words out of the tens of thousands in a language, you can probably express yourself adequately in most situations. These words would be, for the most part, the words that occur with highest frequency in the language. One problem with this approach is that, while the trainee can limit his own speech to the 2,000 words he knows, he has no control over the native speaker he is conversing with. The native speaker may very well use his entire vocabulary and the trainee will be required to understand it if he is to carry on an intelligible conversation.

Now, we know from experience that such trainees can learn to make fairly good guesses about the words they do not know, but it is very difficult to establish firm objectives regarding the number and the required degree of success at such guessing.

So, we know in advance that our objectives will contain some elements of ambiguity--some factors that will not be well delineated and that will cause some problems for the writers of criterion referenced tests.

Can job and task analysis help us to define the objectives with greater precision? One would hope so, but keeping in mind that it will be quite impossible in any practical period of study to gather enough data for full confidence that all possible conversational content and situations have been collected.

It should be noticed, by the way, that by concentrating on vocabulary alone, we have said nothing so far about grammatical structures, idioms, dialectal variations in pronunciation, and many, many other aspects of a foreign language--all of which should be adequately described in the enabling objectives for an instructional system. Of all the separate problems that these raise for analysis, for the writing of objectives, and design of tests, let me mention just one.

It is important that we identify the types of persons that the student will have conversations with, and the circumstances of the conversation, because languages change significantly from place to place and person to person. We are all aware of the fact that we do not speak exactly the same way in giving a briefing to the commander as we do at a poker game with friends, or with our wives. Linguists therefore talk about the different "registers" of the language, and they have a great deal to say about how the vocabulary changes in different situations. What is not so obvious is that the grammar changes, too, and sometimes quite drastically.

Some of these changes might be called grammatical errors by a language teacher, but linguists sometimes argue whether these are really sub-standard usages or represent quite correct grammatical forms for the specific situations. The English language is said to have five distinctly different registers, and it is probable that all other languages have as many. So it is a legitimate question to ask which of these registers--or degrees of formality--of the language we should teach our students to understand and to use.

Now, some of you will have noticed that, while I have not specifically addressed my remarks to the question of criteria, many of these comments about determining the nature of the objectives themselves have also necessarily touched on how well we expect the student to perform those skills. For many jobs, such as Morse code, we can state the criteria with mathematical precision, that is, with a percentage of permissable error at a stated speed, and so on. It is not entirely meaningful, however, to say that a person can carry on a conversation with only 10% error, especially since it is probably not too far from the truth to say that only about 10% of the usual conversation is of any significance. It becomes fairly important, then, to be able to identify where the errors occur. Unhappily, we have no sure way of determining that.

Well then, what can we do in the way of job and task analysis in the foreseeable future?

First, we must recognize that there probably will be no way to solve all these theoretical and technical problems right away. We must, instead, try to find some practical way of reducing the area of doubt about the language skill requirements or the job and reducing the amount of error in our determination of the vocabulary, grammatical features, etc., that we include in a foreigh-language instructional system. In short, we need some research.

We have therefore contracted with Development and Evaluation Associates of Syracuse, New York, for a research study into some of the problems. This contract was awarded only a few weeks ago, so no answers are yet forthcoming. In fact, the details of the procedure are not yet fully worked out since the details of the procedure are exactly what we want to get out of the project. Howevor, I can sketch the effort in fairly broad strokes.

First of all, we must select a manageable sample of the kinds of jobs we are concerned with. Even though we are excluding the jobs in the security services, there are hundreds of such jobs. Also, we must limit the study to a very few of the nearly thirty languages we teach. So far, we have decided to include Russian and Chinese Mandarin, but have made no other choice yet.

Once we select the sample of jobs and of languages, we know we do not need to do an exhaustive job of task analysis on the entire inventory because all the jobs we are interested in also contain duties that do not require use of a foreign language. Those non-language tasks are not a part of the task analysis challenge for DLI.

Even an interrogator has a number of duties that do not involve the foreign language. We can look at the Soldier's Manual for an Army 96-Charlie and pick out those tasks that must be represented in our terminal skill objectives.

The same thing is true for a naval officer in NATO Headquarters, for an Air Force Staff Advisor in Iran, or a military attache in Argentina, except that no handy guide such as a Soldier's Manual exists.

In each case, we have to determine quite specifically what communication skills the job incumbent must have. Does he have to engage in conversations or not? Does he have to give briefings in the foreign language, or merely attend such briefings and be able to summarize them? Does he read correspondence, or write letters himself? Must he sometimes interview people? If so, what kind of people? Is he required to make translations? If so, what must he translate: letters, technical manuals or directives, political news, or what? Our Job Analysis and Standards Division is now doing a fairly good job of identifying these communication tasks.

But once these skill types are identified, along with some standards of adequacy, we need to find out, in some way, what general and special type of vocabulary is needed, what idioms and technical jargon he will encounter, and what grammatical features of the language are either very common or, even if rare, are critical to performance of the job.

It will be one of the contractor's most important tasks to figure out how to determine which of these linguistic features is necessary, and just how many of them will be sufficient, to prepare the student for the job. There are some serious theoretical problems as well as practical problems to be solved here.

But the contract calls for more than that. When he has determined what he believes is a practical system, the contract requires that he prove that it works. He must then try to apply his method to the particular jobs in the particular languages that have been selected earlier, work out any difficulties that arise in the application, and provide us with a workable method that we can apply across the rest of the jobs and languages for which we must devise instructional systems.

The entire contract effort is expected to take 21 months.

I'd like to conclude with a few remarks about what this may mean to those of you who are not engaged in foreign-language instruction. The Defense Language Institute Foreign Language Center is not the only military organization that has problems of domain definition, and difficulties in applying the principles of job and task analysis to determination of both terminal and enabling objectives. Many of the rest of you face serious problems in developing criterion-referenced tests for the special skills you are concerned with.

For example, the theory of criterion-referenced testing requires that all terminal objectives be represented in the final CRT for an instructional system, but some of you share our problem in that the number of objectives is far too large for inclusion in a CRT of practical length. In such cases, the final CRT can only sample the required behaviors.

It seems likely that some of you also have a problem parallel to ours in that the skill you teach is not the actual content of the MOS or the AFSC or whatever, but a skill <u>through</u> which the job is performed. This is a particularly sticky problem for job analysis and for test validation. Suppose we find, for example, a logistical advisor in Iran who is not performing adequately. How do we determine whether his failure is the result of inadequate proficiency in the language or because he doesn't know his logistical channels well enough?

In the Department of Defense, a foreign language is never learned for its own sake, but in order to do something else. Therefore, unlike other military skills, such as performing maintenance on an engine, on which the individual concentrates wholly, the skill of language is one that must be performed while working toward another purpose, such as the obtaining of information or persuading someone else to do something, or whatever it may be. The language skills must be so thoroughly established that they are habitual and require virtually no conscious attention, and are, furthermore, wholly different in nature from the job to be performed.

If nothing else, then, the research project points up the difficulty of applying routine methods of job and task analysis to skills of this type. At best, it may provide some methodological clues that may be of use to schools with similar problems. If so, I assure you that we will report these to you two years from now.

STUDY OF TASK DIFFICULTY USING FIELD TEAMS & THE AFHRL TASK ANCHORED SCALES

Mr. Fred L. Hart

ABSTRACT

The history of the scientific study of work is summarized from early work by Charles Babbage, better known for his early attempts to build a digital computer. Historical work by F.W. Taylor, the Gilbreths, L.H.C. Tippett, H. Munsterberg, and R. Gagne and R. Mager are compared in relation to the task inventory method. The state of the art is represented by the work of Ray Christal at AFHRL and his counterparts in other services. Current capabilities of CODAP, task inventory methods, and the 4-factor model are summarized as a basis for other panelist presentations.

This panel is concerned with Job Task Analysis and with the things currently being done in the several US Services to improve our ability to analyze and describe jobs. Today we will be primarily interested in describing jobs for the purpose of deciding the content of training. We will be talking principally in terms of job task inventories, which have been developed using the CODAP system. Each of the armed services is now able to describe certain enlisted specialties with lists of specific tasks which people actually do on the job. The CODAP system (NOTAP in the Navy) clearly represents the best means yet available for the scientific description of work.

It may help us to recall, for a moment, the history of this effort. Modern efforts to describe work scientifically and to use that data in scientific management are about 150 years old. During that time there has been an evolution of ideas and of techniques which is interesting to recall. Furthermore, it may be useful for us to be aware of the key approaches taken in the field, and of some present alternatives to the task inventory method.

We are talking about description of the work process. Possibly the first scientific attempt in this direction was that of Charles Babbage in England, who in 1832 published his <u>On the Economy of Machinery and Manufacturers</u>. Some of you will remember the name. Babbage was the distinguished mathematician who is the true father of computing machinery. With the primitive means at disposal he built "analytic engines" which computed mathematical tables still used until recently. His most ambitious machine was never built, but would have been equal to a small modern computer. His computers incorporated most features of modern machines, including punched card input, separate registers for data and program, and automatic printout. Babbage is less well known for the fact that he pioneered in the study of the industrial work process and in the rationalization of manufacturing. His methods were reminiscent of "operations research," as that term was used by the British in the '40s and '50s. He was particularly interested in quantifying work, and his practice was in some respects very modern. His book was republished several times in Britain and the United States, and his methods were applied, for instance, to the manufacture of pins and in the British Post Office. As a result, his research in the work process achieved more practical application than his "analytical engines."

Often a genius seems to be before his time; so it was with Babbage. His notions on the computer and on the study of work were not systematically followed up, although they may have stimulated some similar writings by J.R. Perronett in France.

The Engineering Management Approach

It was not until a half century later that F.W. Taylor laid the foundation for a science of work management. Taylor described the basic principles of "time study" by which any job was divided into elementary tasks, measured for time, and ordered for productivity. Taylor's emphasis was upon time for performance. He promoted the production of more products per worker and per unit time. For this reason, some of his results have been considered dehumanizing. He observed motion and decision making only incidentally.

The concept of difficulty is very important to us, because it affects the differentiation of workers on the basis of aptitude. Difficulty Taylor did not address, except as it may have been implied in his use of terms such as "delay," "complexity," "skill level" and "effort." Although he did consider training and worker selection, he did not treat difficulty as an independent concept.

In the 1900s Frank and Lillian Gilbreth originated "motion study," incorporating a consideration for the physiological and psychological capacities of the worker. They used photographs to document work processes, and they implied difficulty when they discussed the problems of learning, fatigue, monotony, attention and decision making. During World War I, the Gilbreths developed training in the assembly and disassembly of weapons, initiating (along with Hugo Munsterberg) the tradition of military human factors research which we are continuing.

The Applied Psychological Approach

Babbage, Taylor and the Gilbreths represent an approach centered on engineering and management. In 1913, Hugo Munsterberg became the first formally trained psychologist to study industrial management, and treated problems of task learning, adjustment to physical conditions and economy of motion. He was active in applied military research during World War I. It is of interest that the practice of testing for personnel selection dates effectively from the Army Alpha and Beta tests that were used during the 1915-18 mobilization.

After the war, L.H.C. Tippett studied the British textile industry, using statistical methods (Tippett, 1935). Tippett's contribution included the concept of activity sampling (Barnes, 1956). The statistical approaches Tippett used foreshadow the methods of modern manpower research, especially in the military, where statistical treatment makes it possible to generalize about the characteristics of

work and of populations of workers. Like others before and since, Tippett avoided confronting the concept of "difficulty" as a task variable.

Time and Motion Study

From these and other efforts a descriptive science of work had emerged by the 1950s, often referred to as "time and motion study." In this classical practice, "job difficulty" was seen as that which forces a worker to perform at less than a "standard pace" theoretically achievable. These ergonomic difficulties encompassed a number of factors which might impede the pace of work, such as weight to be overcome, distances to be moved, hand-eye coordination and asymmetry of movement. Not included was any notion of difficulty resulting from complexity of information processing, nor from effort to acquire knowledge or skill. Central nervous system processing steps such as "search," "plan" and "select" were considered, but were seen only as "...elements that tend to retard accomplishment..." As a result, a standard text on Work Measurement (Abruzzi, 1956) devotes half a column of its index to "delay," but does not list "difficulty." The research so far described laid the basis for a science of task analysis, but it provided no metric for difficulty which would assist the decision to train, to select job content, or to select personnel.

Job Evaluation

Since about 1920 there has been continuing development of schemes for the classification and grading of jobs. That effort has had little interaction with the engineering-management approach or applied psychological approach we have just outlined. Job evaluation is concerned with the practical considerations of wage and grade administration. It seeks to rate jobs on the basis of relative scarcity of qualified workers, or expedient dollar costs for work. The job evaluation matrix is value; "difficulty" is implied, but is not rigorously defined. Task structure is never described, even on a sampling basis.

Systems were first developed which, in general, measured jobs in their entirety rather than task by task. Early techniques used simple whole-job rankings, in which a panel would rank a set of existing jobs. About 1922, the Grading or Classification Method appeared, pioneered by the Bureau of Personnel Research at the Carnegie Institute. This method required the design of a scale, containing descriptive statements of the levels of duty, responsibility of knowledge required at each rating level. Raters would then match jobs to the scale,

occasionally using a factor weighting system, for instance 60% filing, 40% typing. This system survives in the current practice of the U.S. Civil Service Commission. A later system described by Lott required identifying "factors," often a psychologically mixed bag: skill, strength, effort, responsibility, mental requirements, or aptitude for learning might be considered together in a list of 8-15 or more factors. Different jobs were then classified according to the degree or level to which each factor was required for each job. Factors were weighted for value, and job ratings derived by various algorithmic procedures.

A further elaboration was the factor-comparison method, attributed to Benge or to Mitten. This system used a larger number of factors, not common to all jobs. Each factor was ranked independently, using a benchmark job factor scale which might include a dollar rate of pay index to assist factor-to-factor comparability. In spite of the number of "factors," there was no systematic task breakdown.

The use of "factors" in job evaluation is comparable to a similar use of factors in describing tasks. The various CODAP task inventories match individual tasks to numerical values for factors such as difficulty, criticality, frequency of performance and time required to perform. Notice the difference: job evaluation assigns factors to the job as a whole. The task inventory approach describes each task separately.

Job evaluation schemes met the need of organizations to establish hierarchies of jobs which were related to job market conditions, and which could be defended as objective. They contribute little else to management. The concept of "difficulty" they assume is a mix of factors, including factors of aptitude, learned competencies, job attractiveness, and conditions of the labor market.

Modern Developments

During and following World War II, new directions of effort emerged. One was "Operations Research," a British term for mathematical modeling, to find optimum strategies or paths within systems with several interacting variables. From this concept of whole-system quantification came development of the "systems approach" for planning within U.S. Armed Services. The term had at least two implications of interest here:

First, it made obvious the need, in developing new weapons systems, to recognize human requirements and to have trained manpower ready to man the new hardware when it entered service. This required a means to predict the content of future jobs. Manpower research, hitherto a measurement science, was required to undertake prediction.

A second and related implication of the systems approach was that a whole-system model should be applied to the cycle of recruitment, selection, training, and career progression. Scientifically quantified means were needed, by which to make management decisions regarding the apportionment of recruit talent among competing fields, the content of training, when and how to train, how to apportion tasks among workers, and how to structure specialties and skill levels among servicemen. It was to meet this need that the USAF began personnel and manpower research in the early 1950s, and initiated its occupational survey programs.

Meanwhile new dimensions were added to the practice of work study by behavioral scientists.

The term "human engineering" became common. Generally that term identified the measurement of human strength, speed, perception and motor skill in relation to the control of machines. This work contributes to our understanding of the possible dimensions of difficulty. Ryan treats difficulty as task effort required, a function of foot-pounds of physical work, and metabolism. Required effort may be increased by the intrusion of fatigue, physiological stress, or induced inefficiency caused by psychological stress or boredom. Mundell recognizes six categories of physiological "difficulty:" (1) the percent of total body required to perform a task, (2) whether or not foot controls are required, (3) whether hand motions required are simple or complex, (4) level of hand-eye coordination required, (5) grossness of hand motion tolerated, and (6) weight (force) to be moved. Such studies provide a limited taxonomy of physiological difficulty, but are little help where the primary obstacles to task performance are the need for skills or knowledge. Mention of human information processing is still noticeably absent.

The term "job analysis" describes varying methods for acquiring data descriptive of a job. That practice is an advance in relation to both time and motion study or job evaluation, combining features of both. Job analysis seeks to provide a critical job-content profile, using as source data interviews, activity logs, observation logs, critical incident reports, biographical inventories, standardized

tests, or similar material. A profile of required aptitudes, traits and competencies is derived, and ideally is then validated to the population for which it is designed, using a regression equation or a multiple cut-off strategy.

The term "task analysis" often is used interchangeably with "job analysis." That is unfortunate, for there is a critical difference. In Task Analysis the focus is on the component tasks of a job. Tasks are to be described in terms of specific stimuli, standards, conditions and actions, or alternatively cues, conditions, actions, and products. There is an obvious parallel with Mager's statement of the elements required for an instructional objective. Most task analyses have the remaining weakness that they are descriptive, rather than quantitative, and therefore difficult to compare one to another. They resist computer handling, and are not convertible by an algorithm either to training objectives or to aptitude requirements.

Such quantification is possible however, using job inventories as developed by AFHRL.

USAF Research

USAF established its Occupational Research Project in 1958, initiating a systematic program for the study of military and civilian jobs in relation to human resources. A major product of that effort was the development of job descriptive information, under management of the CODAP analysis system, with quantified factor data. AFHRL selected the job inventory approach as being feasible, and as providing data which could be quantified and manipulated by computer. A job inventory is a list of all tasks normally performed by a worker, derived from a survey of job incumbents. AFHRL prepared comprehensive lists of task statements, derived from several sources. After several steps of technical editing and review, those lists were mailed to workers. Each worker checked off those tasks he normally performed, wrote in any tasks which were not listed, and recorded time spent on a relative scale. Using this procedure, the job of any worker or group of workers can be defined by a subset of tasks from the inventory.

As you know, CODAP provides a sophisticated and flexible means for managing the task inventories and associating them with factor data of various kinds. The CODAP system has been adopted or modified by each of the Armed Services for its own use. I will leave it to the panel members present today to comment further on what those services are doing. But I do want to make one point about task inventories and training.

Data for Training

One of the primary reasons for the scientific study of work is so that we can train workers effectively. The systems model for Training - ISD - presumes that training will be based on the content of jobs. The ISD process requires that training development begin with a task analysis. As a matter of fact, the ISD documents are a little vague about how this will be done.

Conventional task analysis was descriptive. As usually practiced, it was useful for management, but provided little specific data for training developers. A typical task analysis reached a level of specificity close to that of a course description, but not approaching the time-by-task by task description needed to write learning materials. A specific job might be described differently by two different analysts, and that data was not suitable for computer management.

The CODAP task inventories take us an order of magnitude closer to what training developers need. These data can be replicated and machine processed. The task statements are moderately specific and use a disciplined vocabulary. Each task can be tagged with valuable data such as the frequency of task performance. Still the CODAP task statements are not readily convertible to training objectives.

A typical USAF task inventory contains from 200 to 1000 or more task statements, describing the work in one Air Force specialty (AFS). Yet a single task statement such as: "Perform alignment of aircraft H.F. Radio Receiver" (AFSC 32850), actually describes over 100 discrete task steps. Each of these steps includes its own distinctive actions, conditions and standards. Just how those atomic learning objectives are to be scientifically described, and based on the objective study of work, ISD does not tell us.

Several of the speakers today are taking action, within their own services, to develop methods by which training may be objectively based on work, at the level of a subtask or the level of a programmed instruction frame. With you, I look forward to hearing the speakers who follow.

A MANAGEMENT FEEDBACK SYSTEM

FOR THE AIR FORCE

MAJ JON L. GROSS

ABSTRACT

Systematic feedback on the management of people is required to improve both management and organizational effectiveness. Rensis Likert's questionnaire, "Profile of Organizational Characteristics," is a proven technique for providing systematic feedback. The military manager's use of this questionnaire will help to improve the management of the most critical resource--people.

A MANAGEMENT FEEDBACK SYSTEM FOR THE AIR FORCE

Most Air Force managers are familiar with the functions of management: planning, organizing, directing, controlling, and coordinating. All managers engage in these functions to some extent in their efforts to improve and maintain the efficiency of their operations or programs. Likewise, to determine their effectiveness as managers, they need some means of measuring the results of their actions on people and on their operations. Rensis Likert's "Profile of Organizational Characteristics" is a proven technique for evaluating management style and organizational effectiveness. Use of the Likert questionnaire can help the manager to determine his current management style and identify areas that should be changed.

A key to management improvement is a positive attitude on the manager's part regarding feedback and a willingness to evaluate the impact of his actions on people and his organization. Such an attitude can lead him to consider various styles and approaches in his attempts to improve his ability to manage. The Likert questionnaire can serve as an initial step by helping him to determine his current management style and by identifying areas that should be changed.

Purpose of Management Feedback System

The participative management approach is a current trend in the management of people in the military. Some managers resist changing from the traditional authoritarian approach for fear that discipline will diminish and people will no longer respond to orders.[1] Most management authorities believe that the participation of people in decisions that affect what and how they do on their job is the most valid, long-term approach to improving the effectiveness of people. Current problems require the fullest use of people in order to be prepared for war if it comes.

Traditionally, managers receive information on how to improve cost control, make better decisions, and manage people. In each of the first two areas, feedback techniques have been developed and incorporated in the management of programs and operations. By identifying areas that require corrective action, the information feedback loop

has proved helpful in controlling programs. However, few techniques are currently used to provide systematic feedback to military managers on what is happening with people as a result of management decisions or actions. Admittedly, "open door" policies and various councils provide some information, but managers normally depend on these approaches to identify and resolve grievances. If things are "OK" or if the level of complaints is reasonably low, management usually takes no action.[2] When these feedback systems identify areas that require change, the manager may focus only on the most visible and apparent problems. Apparent problems may be solved, but the covert cause of a problem may remain to create other apparent problems.[3]

A factor that generates a need for an effective management feedback system in the military is the trend toward longer assignment tours for managers and their personnel. A traditional attitude has been expressed in these words: "Don't let the personality problems bother you and affect your actions. One of you will be transferred in a year or so, and you can get by for that long." Longer tours place greater emphasis on solutions to problems that affect the performance of people. People want to feel that they are needed and useful, that their work is important, and that they have a voice in what happens to them. If they cannot find satisfaction in their jobs and if they cannot expect transfers, they will look outside their jobs for satisfaction. People need to understand the importance of their efforts, know that they have a stake in what happens to them, and feel dedicated and satisfied.

Just as people are keys to greater output with diminishing resources, the manager is the key to more effective use of people. Feedback to the manager not only provides direct benefits in terms of dollar costs; it also informs the manager concerning his effectiveness in the management of people.

> Executives and managers need feedback on their performance and effectiveness, just as subordinates do. One kind of feedback that can be extremely helpful is a survey of employees' perceptions of their supervision, working conditions, opportunities, and other factors. This device is not only communicative in itself; it also is a catalyst to increase the quantity and quality of communications on a continuing basis.[4]

A Management Feedback System

Rensis Likert's questionnaire is a tested measurement tool for providing management feedback on the following organizational characteristics:

- o Leadership processes
- o Motivational forces
- o The communication process
- o The interaction-influence process
- o The decision-making process
- o Goal-setting or ordering
- o Control processes
- o Performance goals and training

The questionnaire consists of 51 questions that measure the perceptions of members about their organization; it does not measure what is good or bad and right or wrong.[5] The manager must determine the characteristics that are appropriate for his organization. The organization does not exist to improve itself; it exists to fulfill a need. Thus, by measuring the characteristics of his organization, the manager can guide it toward the optimum conditions necessary for maximum performance of the mission.

Figure 1 depicts the Likert approach to organizational improvement.[6] To measure the "state" or current condition of the management system, Likert defined three sets of variables as follows:

a. "Causal" variables are independent variables that determine the course of organizational development. These are the only variables that management can change.

b. "Intervening" variables reflect the internal state and health of the organization.

c. "End-result" variables are dependent variables that measure achievements of the organization.[7]

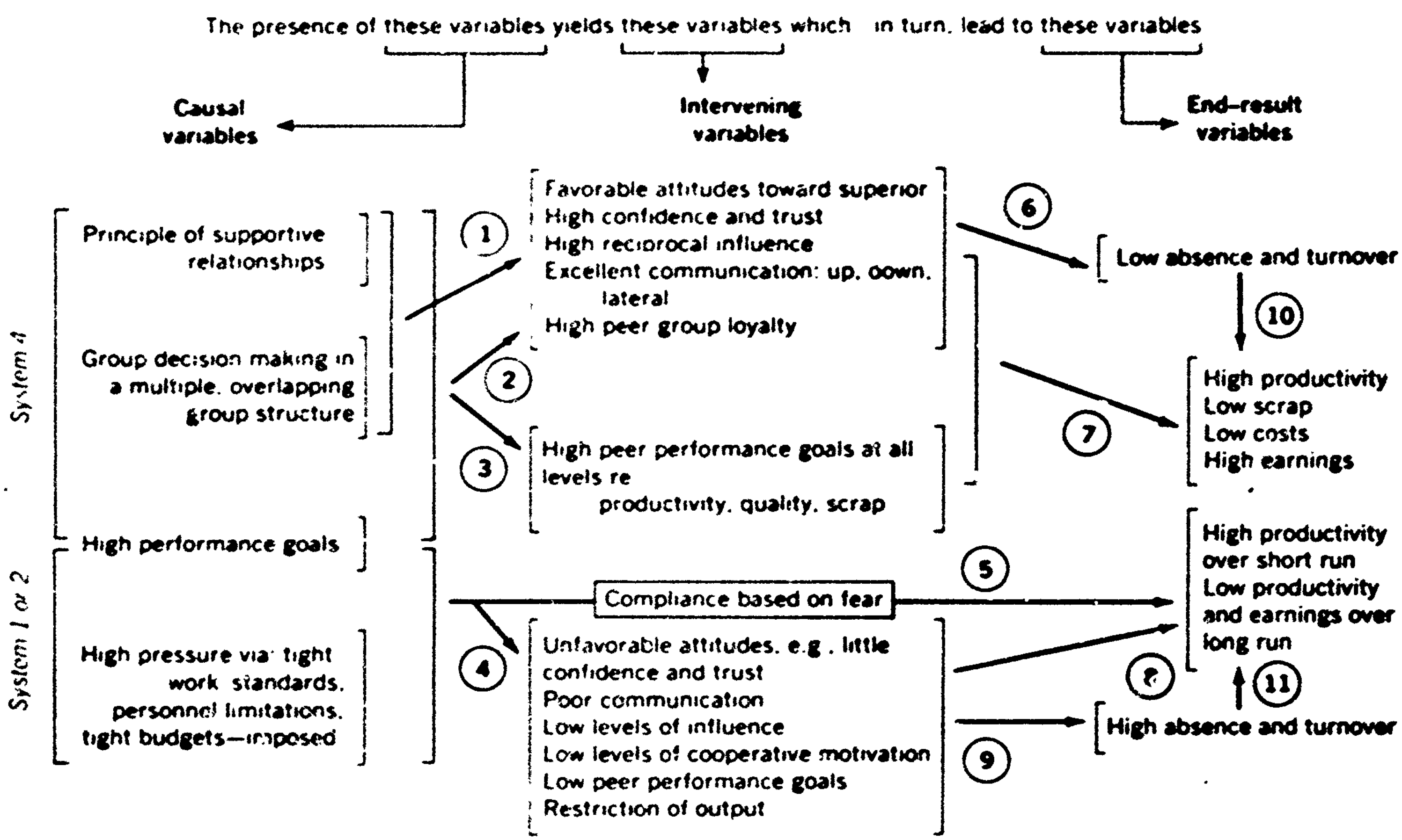


Figure 1. Simplified Diagram of Relationships Among Variables for System 1 or 2 and System 4 Operation.

(From The Human Organization by Rensis Likert. Copyright 1967 by McGraw Hill, Inc. Used with permission of McGraw-Hill Book Company.)

"Causal" variables controlled by management produce changes in the "intervening" variables, which, in turn, change the "end-result" variables. Therefore, the most effective way to change the organization is to modify the "causal" variables.[8]

The Likert questionnaire measures causal and intervening variables.[9] In using the questionnaire, the manager can locate the management system of his organization on a continuum scale ranging from Exploitive-Authoritative (System 1) to Participative Group (System 4). The two remaining categories on the scale are Benevolent-Authoritative (System 2) and Consultative (System 3). Likert states that "the four different systems really blend into one another and make one continuum with many intermediate patterns."[10]

After the manager initially locates the management system of his organization on the scale, he can then, through periodic use of the scale, track the movement of his organization. The thrust of Likert's work is that the organization improves in effectiveness of production and management of people as it moves toward System 4 (Participative Group).

Military Examples

The U.S. Navy uses a modified version of the Likert approach to provide feedback information to Navy managers and supervisors. With the assistance of the Institute for Social Research, University of Michigan, it has expanded the Likert questionnaire and adapted it to Navy organization and terminology. The Navy program, known as the Human Resources Management Support System, includes three major categories:

a. Human Resources Management--leadership, management, and overseas diplomacy.

b. Equal Opportunity/Race Relations.

c. Drug Abuse Control and Alcoholism Prevention.[11]

This discussion limits the Likert approach to the leadership and management areas, which the manager directly controls and influences.

Figure 2 provides one example of the data obtained by the Navy's Human Resources Management Survey.[12] A comparison of Figure 2 with Figure 3 shows the similarity between

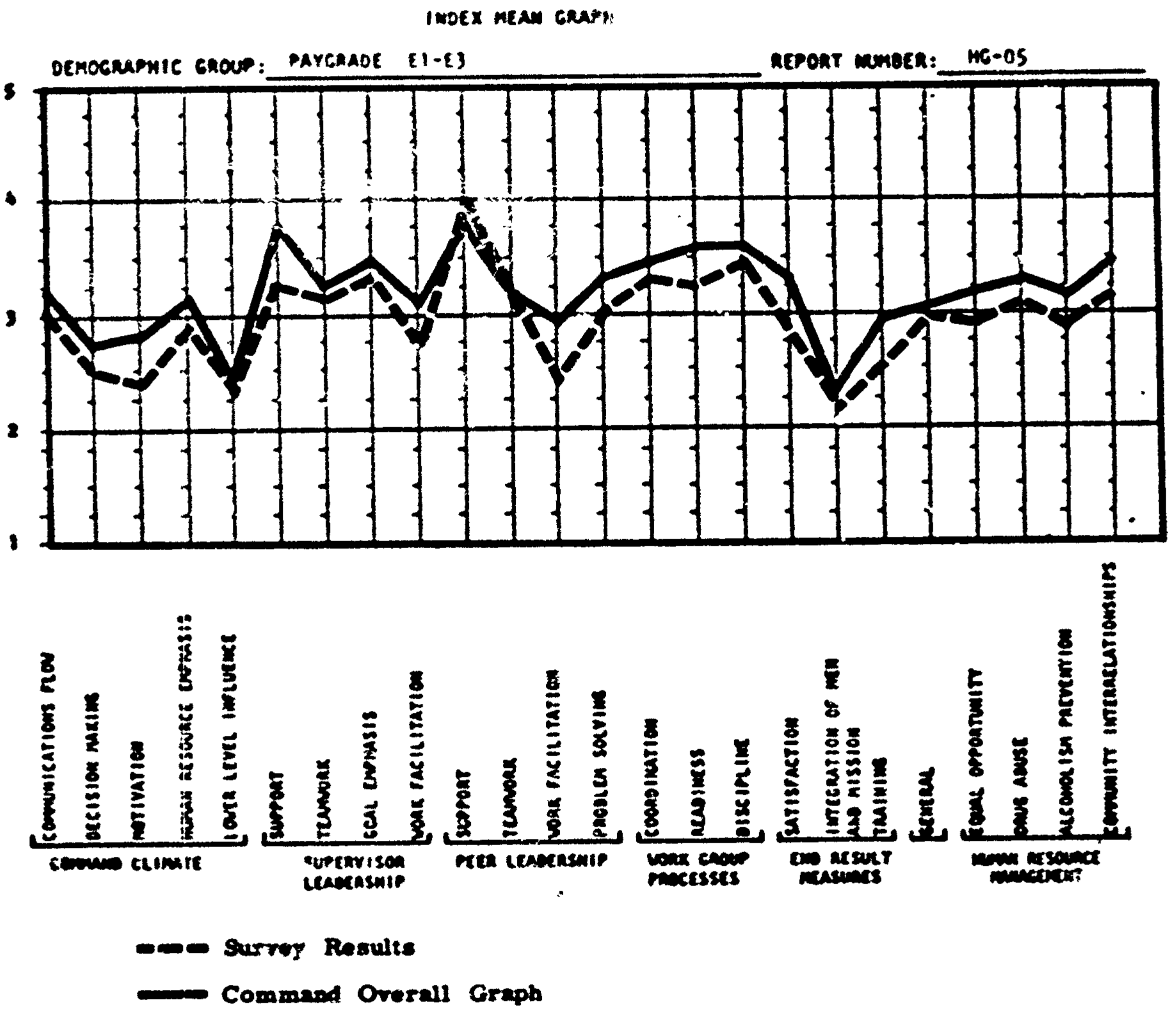


Figure 2. Examples of Navy Survey

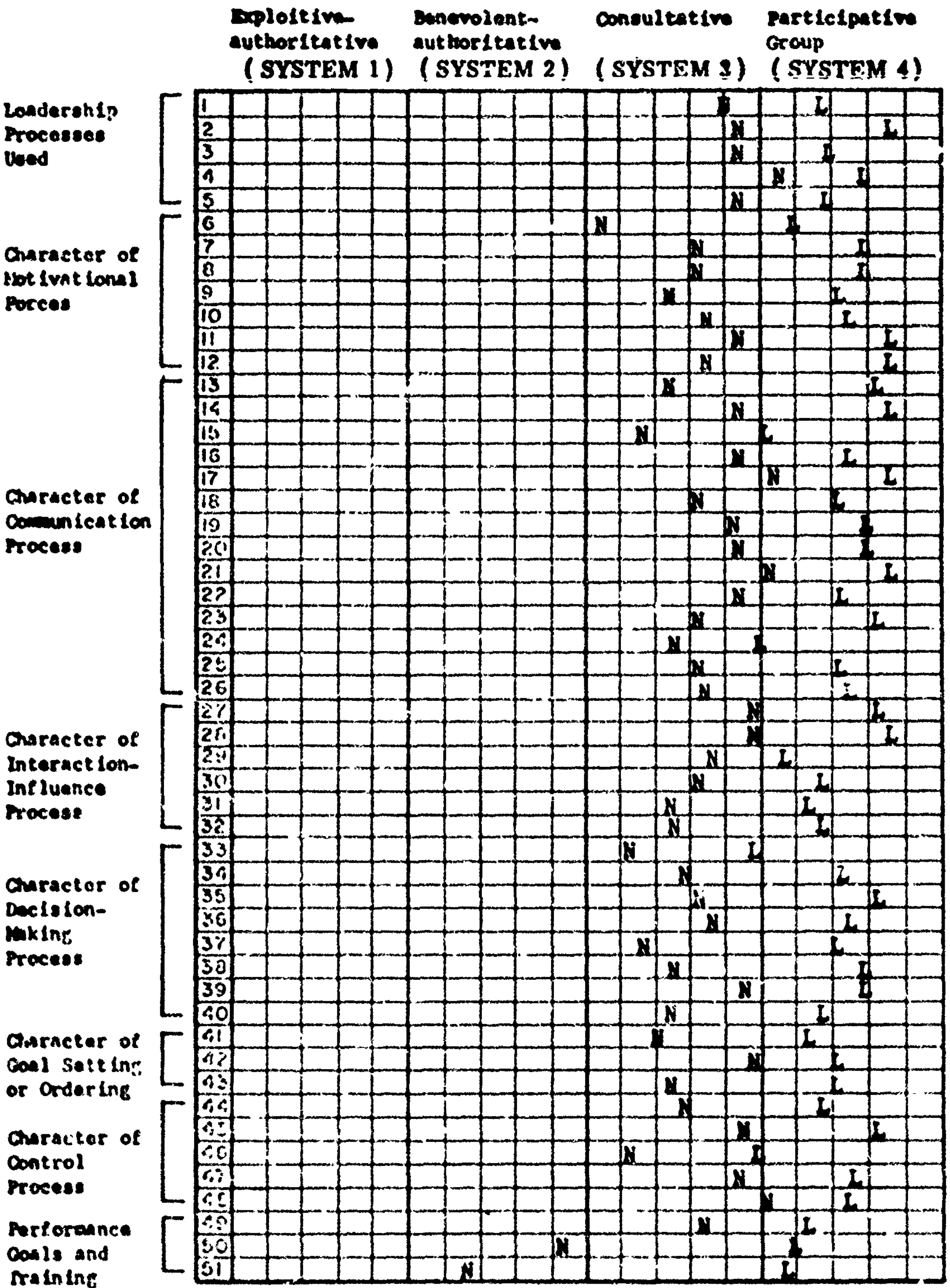


Figure 3. Organization Average

the results achieved by the Navy survey and the results from the Likert questionnaire. The broken line represents the survey results; the solid line represents the Command Overall graph. The Navy approach allows the manager to compare average perceptions within his organization with average perceptions for the command. The principal difference between the Navy questionnaire and the Likert questionnaire is that the members did not indicate their organizational preferences.

A potential benefit of the Navy program is a determination of the effectiveness of certain management styles as they apply to specific types of military units. Managers with appropriate styles can then be assigned to jobs that match their individual styles. Another benefit is that the manager can monitor the trend of organizational change and the relative magnitude of the change over a given period.

To date, the Air Force has made only limited use of similar data-producing instruments to improve management effectiveness. The Leadership Management Development Center at Air University has used surveys in specific management problem areas after the problems have been identified through interviews.[13] The survey results cover only short periods; the long-term effects of management decisions have not yet been targets for investigation. Emphasis has been given to immediate problems and immediate results.

The author of this article and two other Air Force officers applied the Likert questionnaire to an Air Force research and development organization.[14] The survey results for the organization are shown in Figure 3, which displays the average value for NOW responses and LIKE responses.[15] NOW responses are measures of members' perception of the organization at the time of the survey. LIKE responses indicate the characteristics that members would prefer in the organization. In addition to the summary results, the survey included specific responses by military rank and civilian grade on each question.

Implementing Survey Feedback System

The survey feedback system should be implemented at organizational levels that allow subject personnel to associate with a specific mission, such as developing a weapon system, maintaining a weapon system, or providing a particular service like transportation. That is, people in small

organizational units can more readily identify with the program and participate in the effects of the program.

Figure 4 shows a diagram of the survey feedback system.[16] The manager must decide whether to use a survey feedback system in his organization, and he must have support or, at least, acceptance of upper-level management. Basically, the proposal is a bottom-up program in which the manager using the system is supported by his superior.[17]

The plan for implementing the feedback system should be developed jointly by the manager and his subordinate supervisors. By including the supervisors in the planning phase, the manager can insure commitment and understanding.

In the same manner, the manager should explain the survey feedback system to the members of the organization. The purpose and use of the survey, presented and discussed with members, should aid in gaining acceptance of the system. The explanation of the survey must stress the fact of respondent anonymity. Emphasis on the anonymity of the respondents should encourage candid responses. As trust in the manager improves, relationships between the manager and the members of his organization should become more mutually supportive.

The time required to implement the system can be shortened with an existing questionnaire of proven validity, since a new survey requires testing for reliability and validity. Problems in the initial phase of the system will probably center around implementation and understanding of the survey questionnaire. Once the system has become functional, feedback for the manager should be forthcoming.

Upon completion of the survey, the manager should tabulate the results and present them to the members of the organization. Sharing of results encourages an open atmosphere and gives members an opportunity to compare their perceptions with group responses. Furthermore, common understanding of the results should enable the members to participate in identifying group problems and developing proposed solutions. By participating in the development of proposed solutions, members will be more committed in implementing the solutions.

With the initial survey results as a starting point, the manager can monitor immediate changes in his organization and, through periodic use of the survey every three to

PREREQUISITE: CLEAR UNDERSTANDING OF WHY THE ORGANIZATION EXISTS AND THE GOALS FOR THE ORGANIZATION.

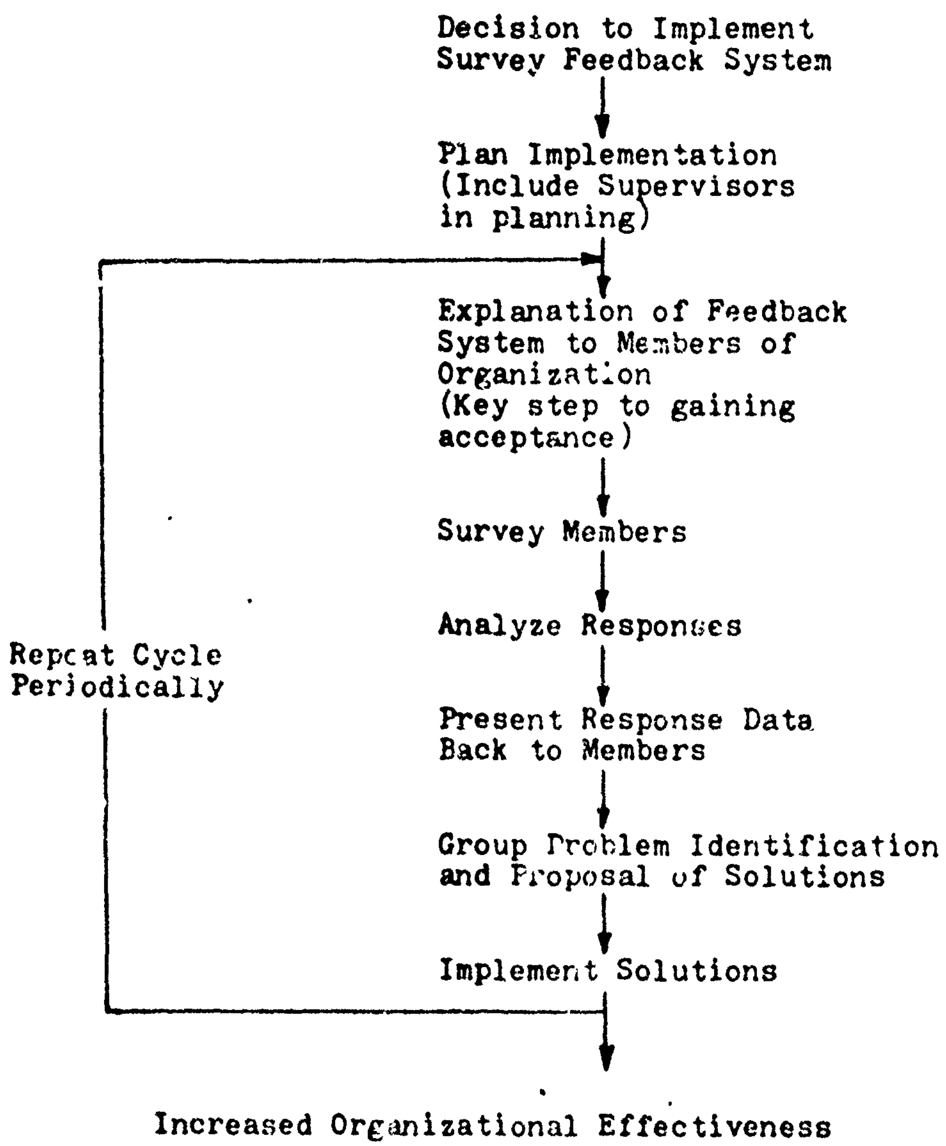


Figure 4. Survey Feedback System

six months, he can evaluate changes in the characteristics of his organization and the direction of the changes. He may find it helpful to have a management consultant; the Leadership and Management Development Center at Air University can provide such support for Air Force managers.

If several units within a large organization use the survey feedback system, the measurement effort can be performed at various levels by applying Likert's linking-pin concept to tie the levels of the larger organization together.[18] It is not necessary, though it is desirable, that all units use the program. One potential problem of using the survey in some, but not all, organizations is that people may be rotated to units that have not implemented management feedback systems.

Summary

This article suggests that the Air Force needs a tool for systematic measurement of management and organizational effectiveness. The results of systematic measurement of management effectiveness should be available to the manager in the same manner as cost accounting and schedule control systems. A management survey is available to improve management of people.

The purpose of a management measuring system is the improvement of management and organizational effectiveness; such a system is not intended as a means of grading the manager's performance. The idea is to close the gaps between action, reaction, and corrective action. This is necessary because management is not an exact science. Action intended to produce a given result can produce something quite different, and the time between reaction and corrective action can be critical.

Consistent with accepted organization development concepts, management actions must be tailored to the organization and its situation. However, tailoring requires measurement of the current state of an organization and establishment of a base for comparing the direction and magnitude of change.[19] Just as the trial and error approach to weapon system acquisition is now an unaffordable luxury, so is the trial and error approach to the management of people. Use of a proven instrument such as the Likert questionnaire can provide systematic, meaningful feedback for improving the management of the Air Force's most critical resource--people.

FOOTNOTES

1. General David C. Jones, USAF, "Discipline and Human Relations," USAFE Pamphlet 30-100 (May 1975), pp. iv-vi.

2. Mary C. Meyer, "Managing Organizational Change," Personnel Journal, Vol. 55, No. 11 (November 1976), p. 571.

3. H. Kent Baker and Ronald H. Gorman, "Diagnosis: Key to O.D. Effectiveness," Personnel Journal, Vol. 55, No. 10 (October 1976), p. 508.

4. John Zenger, "Increasing Productivity: How Behavioral Scientists Can Help," Personnel Journal, Vol. 55, No. 10 (October 1976), p. 515.

5. Rensis Likert, The Human Organization: Its Management and Value (New York: McGraw-Hill Book Company, Inc., 1967), pp. 197-211.

6. Ibid., p. 137.

7. Ibid., pp. 26-29.

8. Ibid., pp. 75-77.

9. Ibid., p. 143.

10. Rensis Likert, New Patterns of Management (New York: McGraw-Hill Book Company, Inc., 1961), p. 234.

11. Department of the Navy, Bureau of Naval Personnel, Commander's Notebook: Human Resource Management Support System (NAVPERS 15279), January 1976, p. A-1.

12. Department of the Navy, Bureau of Naval Personnel, Survey Guided Development: A Systematic Guidebook of Administration, Reduction and Analysis of the Human Resource Management Survey (NAVPERS 15304, Copyright 1976, Organizational Diagnostics Associates), pp. 3-24.

13. Major David A. Wilkerson, Ph.D., USAF, Research Division, Air University Leadership Management Development Center, Interview, January 1977.

14. Ronald D. Bouvier, Jon L. Gross, and Warren Von Worley, "A Management Study of the Behavioral Aspects of a Military Research and Development Organization" (Unpublished research project, University of Southern California, Los Angeles, California, 1971).

15. Ibid., p. 21.

16. The method of implementing the survey feedback system has been adapted from the Navy's Survey-Guided Development: A Consultant Manual for Human Resource Management Specialists (NAVPERS 15264), November 1974, pp. 1-16, and a similar approach presented by Baker and Gorman in their article, "Diagnosis: Key to O.D. Effectiveness," p. 507.

17. Ralph H. Kilmann and Bill McKelvey, "The MAPS Route to Better Organization Design," California Management Review, Vol. XVII, No. 3 (Spring 1975), p. 24.

18. Likert, New Patterns of Management, pp. 113-115.

19. Baker and Gorman, p. 506.

A JOB ENRICHMENT EFFORT FOR THE MARINE CORPS

Lt Col Richard P. Capatosto

ACKNOWLEDGEMENT AND COMMENT

I take no credit for the theory, concepts, or methodology developed in this paper only their application to Marine Corps interests. Motivation-Hygiene Theory and Orthodox Job Enrichment (OJE) strategy belong to Doctor Frederick Herzberg. The OJE office at Hill Air Force Base Air Logistics Command was extremely helpful in providing me their personalized methodology, advice and project results. I am indebted to both these sources with whom I have worked and collaborated for nearly three years. I hope I have not misused their knowledge and wisdom.

This is a concept paper not a Marine Corps position. It is presented in the hope that I may derive valuable feedback from learned conferees. Subsequent to the conference, I shall present the essence of this paper to proper authorities in the Marine Corps with a recommendation that we TRY IT.

Again, my only claim in this effort is its organization and the hope that through it I may arouse the interest and positive actions in the Marine Corps. Total credit for the substance of this paper is credited to Doctor Herzberg's writings and teaching; the major works of which I have referenced at the end.

ACKNOWLEDGEMENT AND COMMENT

I take no credit for the theory, concepts or methodology developed in this paper only their application to Marine Corps interests. Motivation-Hygiene Theory and Orthodox Job Enrichment (OJE) strategy belong to Doctor Frederick Herzberg. The OJE office at Hill Air Force Base Air Logistics Command was extremely helpful in providing me their personalized methodology, advice and project results. I am indebted to both these sources with whom I have worked and collaborated for nearly three years. I hope I have not misused their knowledge and wisdom.

This is a concept paper not a Marine Corps position. It is presented in the hope that I may derive valuable feedback from learned conferees. Subsequent to the conference, I shall present the essence of this paper to proper authorities in the Marine Corps with a recommendation that we TRY IT.

Again, my only claim in this effort is its organization and the hope that through it I may arouse the interest and positive actions in the Marine Corps. Total credit for the substance of this paper is credited to Doctor Herzberg's writings and teaching; the major works of which I have referenced at the end.

JOB ENRICHMENT IN THE MARINE CORPS
(A CONCEPT)

BACKGROUND

As Defense Department funds become tighter and the pressure increases to do more with existing resources, we must find ways to improve the performance/productivity of our manpower assets. This fact of life is exacerbated by the shrinking manpower pool and generally expanding economy which will reach its full impact on recruiting in the early 1980s. If we are to compete successfully with the other services and the civilian market for the high quality people needed in the Marine Corps, we must seek new often bold initiatives to attract, train and retain Marines.

During the first half of this decade, the Marine Corps experienced an alarming increase in unauthorized absence (UA) and desertion rates and a decline in retention rates. Lack of quality Marines was considered to be the most significant reason for these unfavorable trends. In 1975, the Commandant of the Marine Corps established recruiting, discharge, and retention policies designed to raise the overall quality of Marines in the Marine Corps. Consequently, the unfavorable trends in UA, desertion, and to a lesser degree retention, have been significantly reduced.

Unfortunately, these very successes beget problems of their own in that more "highly-educated" Marines require challenging jobs, jobs with substance that draw on their abilities, jobs with which they can identify and within which

they can grow. In short, high quality Marines require the highest quality of leadership and manpower utilization if we wish to attract and then both sustain them and retain them.

Major efforts continue to be exerted relative to the best methods for effectively classifying and assigning Marines. That effort of balancing individual ability, potential and desires, to mission requirements must receive continuing attention for it is the basis upon which we can build meaningful jobs and a motivated force of Marines.

Assuming the classification effort is optimized, considering current constraints, we must then take cognizance of the following chronology of issues:

First, todays Marines, while possessing the same basic needs as always, also brings to our leadership picture a significantly new frame of reference and a new set of contemporary expectations which must be met if we are to employ their talents optimally in the accomplishment of our mission.

This issue of contemporary expectations deserves our utmost attention. Possibly no one would deny any of our time tested leadership principles and the identified traits of successful leaders. Neither should we deny the social and psychological changes manifested in the youth of today. Tradition has long been a mainstay of military life. It provides each of us with an identify; it provides us with a code of behavior. Nonetheless, tradition must also be kept contemporary. Failure to do so will evidence an inability to identify with the organization

in daily, ongoing processes. This fact alone requires that we look at our recruits "as they are" and "not as we want them to be". Once we get on a common, humanistic, contemporary plane, we can then begin to blend more closely, personal and organizational goals into a common mission i.e. we must build on their talents. Doctor Charles Moskos, Northwestern University at an attrition conference in Washington, D. C. in June 1977, admirably identified the salient differences in attitudes between the typical draftee in the late 1950s and early 1960s with those of the typical volunteer of the late 1970s.

INSERT CHART 1 ABOUT HERE

Doctor Moskos stated that the above topology should direct attention to the kinds of expectations and behavior of the new volunteer which can lead to high levels of attrition if not properly handled. In short, pride and motivation are concepts that must be kept contemporary. If Doctor Moskos' analogy has merit, and I believe it does, then the recruit of today wants to know why and how todays situations are relevant to him. Yesterdays' heroes and successes were important in the derivation of principles but we must now show their relevance in an ever-changing society and in the workplace that he confronts daily. Again, this issue of contemporary expectations is an area requiring much exploration and research.

Second, recruitment offers the young potential Marine, a challenge to draw upon the full range of his talents and then to improve his competence in order to sustain himself as a

flexible, total Marine. We infer he will become a totally competent, balanced professional.

Third, recruitment, in effect, establishes a psychological contract between the potential recruit and the job we offer him in the Corps. It is here that the expectations of the potential Marine are established. Those expectations and concomitant challenge should then be heightened in Recruit Training where his professional training is initiated and the often arduous process of conversion from civilian to self disciplined Marine is to be accomplished.

Fourth, upon reporting for on the job training (OJT) and the normalizing of a Marine's work day, he finally comes face to face with the reality of his job. His identity as an important (or unimportant to him) individual in a larger organization with a mission is established here. His niche in the unit solidifies at this point. His relation to his job and his peers for the forseeable future develop here. The fulfillment of his psychological contract with the Corps, which was established during the recruiting process, should come to fruition at this point in the Marine's enlistment. If he does not develop positive attitudes as a result of work that is meaningful to him as well as the Corps, then the gap between expectations and reality will produce either malbehavior or at least nondescript behavior--neither of which is beneficial to either party to the contract. Job design is paramount at this point. Sick jobs rapidly initiate feelings of uselessness and disillusionment for at this crucial point--his job actual--the Marine sees what is truth to him--usefulness of his talents or the converse.

To reduce or eliminate the gap between expectations and reality requires a thorough understanding of two issues: (1) what we do TO our Marines (context/environment/extrinsic factors and (2) what we do WITH our Marines (job content factors; factors that provide meaning to what people do, motivating factors; in short how we USE our Marines). These two issues are manifestly important to the mental health and psychological growth issues so paramonunt to the young, high quality recruits we wish to enlist in our Corps. Referring to Doctor Moskos' topology again, we can plainly see that in adequately answering the questions of how we treat and how we use our personnel assets, we begin to answer the apparent existential question implied in the profile of the new volunteer, which is to bring meaning and purpose to the quality of each Marines' personal and work life.

THE ISSUE

It is postulated that if we are seeking an increase in motivated behavior and if basic propriety in the classification/ assignment of Marines (competence to perform) can be assumed, then a program of job enrichment should materially benefit an organization.

Job redesign requires knowledge of the elements/indicies of good and bad jobs, a knowledge of situational dynamics and ability to apply that knowledge in the existing workplace. In short, we need to operate from a theoretical framework from which we can derive principles of sound job redesign. Such a program must ultimately support the Marine Corps proposition that

responsibility for command performance rests with the leader of the command. Accordingly, any program effort must, in the final analysis, be employed by those who lead--not some external agent as is the case in most organizational development strategies.

Doctor Frederick Herzberg of the University of Utah is the creator of Motivation-Hygiene Theory (M-H-T) which provides a comprehensive framework for his Orthodox Job Enrichment (OJE) program. I content that this theory and job enrichment philosophy constitute both a managerial strategy and a process conducive to success. It is a strategy that involves theory, concepts and views on the nature of man and a direction for management. As a strategy it incorporates the dynamics of human relations, communications, administration; salary etc.. As a process it involves putting theory to practical use via the institution of good job ingredients into the work place.

It would be ludicrous to attempt a comprehensive explanation of Doctor Herzberg's theory and applications in this paper. However, the salient points of the theory are necessarily reviewed since they provide the theoretical underpinning for understanding behavior and attitudes, situational dynamics and OJE methodology.

STATEMENT OF M-H-THEORY

If any theory is to be consistently applied and tested, then its premises must be assumed correct. In M-H-Theory one must assume the bi-dimensionality underlying the feelings of satisfaction and dissatisfaction i.e. the dual nature of man.

The theoretical basis for job enrichment begins with this particular view of the nature of man. Though man exists at all times as a unity, conceptually we view him as having two distinct natures. Each nature has its accompanying need system. One nature--his purely biological nature is similar to that of other animals--an overriding concern to avoid hurt, discomfort, dissatisfaction from the environment e.g. the work around him. This avoidance behavior pertains to both physical and psychological situations (cold, heat, danger, loneliness, feelings of inferiority etc.). By avoiding environmental situations that can cause him discomfort/pain/hurt, he does not achieve positive meaning in his life. He merely avoids discomfort. Because man is so high on the phylogenetic structure, he possesses an almost infinite capacity to learn and experience. Accordingly, the number of environmental factors that can cause him hurt come in an infinite variety and in an equally infinite number of shades e.g. too hot, too cold, too much noise, too quiet etc..

Herzberg called these environmental factors that man is faced with on the job as well as in everyday life "hygiene factors" because when managed properly they prevent dissatisfaction. Other authors have referred to them as maintenance factors. They are extrinsic to the person. They deal with the issue "How well are you treating me?". Their overriding dynamic is avoiding pain/discomfort from the environment. These factors must always be maintained at a reasonable level, for significant deficiencies in any maintenance factor causes pain and like a headache--will impair all physical and psychological processes until near

homeostasis is reached. Only then can one talk about motivated behavior.

The second aspect of man's nature deals with his distinctively human need for continuous psychological growth. Unlike other animals who are primarily preprogrammed in their behavior with a capacity to learn a few new things, man is born very dependent, is nurtured for a protracted period, but who begins learning at birth and can learn new things until biological death. When man does not receive cortical stimulation he begins to display that common human characteristic--boredom. The human side of man then demands approach behavior in a quest for constant psychological growth. He gains from experience. If he doesn't have these growth experiences he is not any worse off--he merely gains nothing--he senses an emptiness or a lack of fulfillment. He feels no satisfaction in this work. Conversely, if he is experiencing new growth, through challenging work, he experiences feelings of personal growth and worth, feelings of satisfaction for he has added to his own personal growth fiber.

The need system associated with the human part of people is served by the elements of the job itself. These factors that lead to growth of people are termed "motivators". These items make the worker want to do his job because it meets his human need to grow psychologically. Meeting these needs is the specific managerial/leadership challenge we face today in the work place. When met, they pay big dividends for an organization as well as the person.

The following graphically displays the two continua inherent in M-H Theory.

Job Dissatisfaction —— Biological Nature ——→ No Job Dissatisfaction

No Job Dissatisfaction ←—— Human Nature —— Job Satisfaction

The following diagram reflects the motivation and hygiene factors associated with the theory (arranged in normal profile format):

INSERT CHART 2 ABOUT HERE

Traditionally, management has placed almost total concern with meeting the maintenance needs of people. This has resulted in people movement but not motivated work. In movement we cause a person to do something he wouldn't ordinarily do by using either a reward or a threat. When someone is motivated to do a job, he does it for something contained in the job--he turns on his own internal generator that doesn't need to be charged by the leader/manager. These motivator needs are not typically met in work today. That is why we need to restructure so many jobs.

Before proceeding to OJE per se, let me summarize the dynamics associated with both hygiene and motivator factors. Such an analogy clearly distinguishes the two continua phenomena.

INSERT CHART 3 ABOUT HERE

<u>OJE ROOTS</u>

OJE operates from a motivated behavior relationship drawn from M-H Theory.

INSERT CHART 4 ABOUT HERE

The first ratio of ability over potential $\frac{(A)}{(P)}$ relates to what a person can do; not what he is like, which is a common displacement error. The more ability a person has, the more he can be motivated to perform. This ratio relates to personnel selection and classification; getting people into the right jobs where they can develop their potential and abilities. Concomitantly, training programs are involved here, to maintain personal development in our rapidly changing technological industry which tends to force rapid obsolescence of talents, techniques, and consequently, people. When a person is in job difficulty, look to his competence first--"can he do the job?". All too often managers are told they must motivate underachievers when in reality, the underachievers lack competence to perform. Motivation therefore becomes a slogan to protect ego against imcompetence.

The second ratio, opportunity over ability $\frac{(O)}{(A)}$, determines how much of the person's talent is allowed to show itself. No one can be motivated to do a good job unless he has a good job to do. Since attitudes are manifested from the behavior of people, we should not expect enthusiastic, motivated attitudes in people who are given fractionated, dull, repetitious, "mickey mouse" jobs. Given a job that uses only a small fraction of a person's ability/potential, a job with little or no challenge or opportunity to grow and to achieve, a job with no place to grow into via recognition for achievement and thus advancement to new responsibility; given a job which denudes a person of the chance to become what he can become, then that person will exhibit behavior and attitudes characteristic of a person in a "sick" job--a job that needs enrichment.

Reinforcement of motivators with more motivators and the reinforcement of hygiene for hygiene purposes is the last factor in the equation. Hygiene must always be kept reasonable. Any significant deficiences must be alleviated for no one can be motivated while he is in a state of hygiene deprivation. Relieve the hurt and then proceed to motivate. On the other hand, when satisfaction is being attained, when people have a good job with motivators present, they must receive reinforcement of those motivators. Does the appraisal system reinforce growth behavior which often involves risk, challenge, a chance to achieve and be recognized, does it offer opportunity to advance and achieve more--to become "what I can become"?

OJE APPROACH

Any attempt to reevaluate a "sick job" requires a recognition of the indicies/characteristics/symptoms of a "bad job" as well as knowledge of the ingredients that should be injected into the rejuvenated job. Briefly, the indicies of a "Bad Job" are:

INSERT CHART 5 ABOUT HERE

Once the indicators of a poor job are recognized, the OJE ingredients of a "good job" can be utilized to enrich the position and improve the motivational aspects of the work.

INSERT CHART 6 ABOUT HERE

Enriched jobs may not have all these ingredients but a conscious infusion of as many as possible of these enriching ingredients will certainly enhance motivated behavior. These ingredients are derived from motivators but are more realistic

and practical to work with when focusing on job elements than the more ambiguous terms of achievement, growth etc. found in the standard profile.

Finally, an operating framework is necessary within which the job enriching motivators can be properly applied; in essence--the OJE work process must be understood. Essentially, OJE is an approach which requires top down understanding and support while the application starts at the grass roots level and works up. Diagrammatically OJE looks like this:

INSERT CHART 7 ABOUT HERE

The job is analyzed with an effort to vertically stack motivators in the job. To do this we "push down" some currently higher responsibilities, to the job; concomitantly we incorporate the job's pre and post processes which fractionate jobs so often (e.g. prepare own preliminary work such as rough drafts for a clerk and authorize that worker to sign and complete the work task) and finally strip away menial, routine tasks wherever possible. As the job is enriched, redefinition of higher level jobs is necessary. These higher jobs are then reorganized toward better management/supervision functions such as planning, training, etc..

OJE ISSUES

As stated, the strategy of Orthodox Job Enrichment is to start at the bottom of the management pyramid and work up. First-line supervisors are involved in enriching the jobs of the line workers. Later, the OJE principles are applied to the jobs of the first-line supervisors, and so on. This

approach is necessary because enrichment of one level has an unavoidable impact on the jobs of the immediate supervisors. Additionally, the bulk of the work force is on the line and this is where the most deflated jobs exist.

A pivotal person in the enrichment process is the keyman. Each organization must have one or more keymen. Keymen are individuals from within the organization who have had extensive training in job enrichment theory and application and have a broad knowledge of the functions and operations of the organization within which they operate. In the OJE process, they are involved extensively at first, but to a decreasing degree as the approach becomes more of an accepted and adopted strategy.

Implementation is on a project basis. Projects are approved by commanders. The first phase in the project is the establishment of implementing and coordinating committees. The implementing committee is comprised of the supervisor of the area to be enriched (the key supervisor) and other first and second level supervisors who can be of assistance in the implementation of job enrichment. Size of the implementing committee varies from four to eight. The coordinating committee is made of middle and upper level management with which the unit involved interfaces. Typically, it consists of four to eight members. The keyman serves as advisor to both groups.

After selection of the committees, the keyman conducts training in Motivation-Hygiene theory. Training usually consists of 30 to 40 hours of theory including classroom exercises to highlight the major points of the instruction.

Following the training effort, the implementing group has the task of enriching the jobs under consideration. The technique used is called greenlighting and redlighting. Greenlighting, or brainstorming, utilizes the concept of deferred judgment. The purpose is to generate as many ideas as possible about how to instill motivators into the jobs. After a list has been green lighted, the group enters a redlight session. Here the ideas are evaluated to determine which will be considered for immediate implementation, which will be kept for possible future implementation, and which will be discarded. A time-phased implementing plan is developed.

The implementation of the accepted greenlight items is the area where the coordinating committee can be most useful. Having been trained in the theory, this committee can now understand the strategy behind the changes suggested by the implementing committee and be of assistance in removing roadblocks to implementation. The coordinating committee also develops the measurement plan for the project. Upper level management, as members of the coordinating committee, provide important assistance to the project by making it possible for the key supervisor to make the changes he has decided are needed. They will also need to make adjustments to their management strategy as more and more of their employees have their jobs enriched.

OJE SUMMARY COMMENTS

When done correctly OJE gives the worker a job to do that provides him with the motivation to work because the job provides a means to meet his human need for psychological growth.

The supervisor/leader is now able to return to doing supervisory functions and get away from babysetting functions. He will be managing workers who have responsible jobs. The leader can differentiate between "things" that prevent dissatisfaction and the "things" that promote satisfaction. Higher levels of management reap the benefits of motivated performance from lower level performers.

OJE is a powerful management strategy to be sure--but it must be kept in perspective. It is not a panacea for management ills; it is not easy to do--it is work; healthy work possibilities must exist i.e. you cannot motivate people to do meaningless work; it does not offer a cookbook solution to the problem of motivation and to work. Finally it must be accepted as a personal strategy by the leader or it is doomed to failure from the onset.

OJE AND THE MARINE CORPS (A PROPOSAL)

Since M-H Theory and OJE should become a way of thinking and a way of leading and not an externally imposed edict to improve this or that index, it seems obvious that four events must take place. First, our top level leadership e.g. the Commandant, DC/S for Manpower, CG of FMFLant/Pac and Division of Wing commanders must sense a need for it and overtly decide to pursue the effort. We must "buy in" all the way or leave it alone. Next, we must gain a sense of direction. What are we after? Third--where do we establish a test project and finally, how do we institutionalize a program.

TOP LEVEL LEADERSHIP

Each of the leadership echelons mentioned above have or will soon be briefed on theory, concepts and methodology. To date, each has shown a positive propensity to proceed with a test project. Only with their executive management support can upward changes be accepted as well as assets dedicated to the effort. Additional briefings will be required if a project is approved.

WHAT ARE WE AFTER

Combat readiness through well trained, highly motivated Marines; professionalism is our goal. Emphasis must be on people. Are we providing realistic, challenging training? Are our people responsive to the training (productivity)? Can we improve their perspective of themselves, their Corps, their mission?

Traditionally we have looked at motivation/morale indicies like U/A and desertion rates, attrition/reenlistment rates, racial incidents, enthusiasm/unit cohesion and most important--how the job is done in the field. These are still a reflection of the unit and they should remain as interim indicies. New indicies may be discovered as any project progresses. Regardless of the index, we are looking for that elusive quality of "attitudinal change" that improves combat readiness. At this point, I do not feel we should look for monetary gains which may well be fallout or even a new index in a project. This could be especially so when we work in technical fields.

Admittedly, some of these sought indicies are "soft" and difficult to measure. Again, my faith in the attitudinal-affective results that are a mainstay of M-H Theory come into play. Go and ask the Marine--he will tell you what no questionnaire can ever do--give you feelings, impressions expressed in a real setting, and he can articulate changes he perceives over tim . I believe this "flavor" in personal perceptions is imperative to the results we hope to achieve.

WHERE TO ESTABLISH A TEST BED

I believe we must build on the basic experience of those who have already tried this approach--the Air Force Air Logistics Command (ALC) effort which has realized such success. Accordingly I am recommending we establish two test projects--one technical project with an aviation squadron and one project in an infantry battalion--our cutting edge--a place most difficult to produce high visibility results but a place where attitudinal changes can mean so much in the execution of mission. Improvements in this latter area will have a tremendous impact on our entire leadership package which must ultimately become the vehicle to institutionalize the entire job satisfaction-OJE effort.

INSTITUTIONALIZATION

Once the effort at the project level is deemed a success, we must then incorporate the methodology into our school system. It may well provide improved boundaries and foundation of our current leadership package which is heavy in human relations activities. This new concept places human relations into true

perspective--a necessary hygenic/maintenance factor (interpersonal relations) within the total leadership sphere. Accordingly, teaching M-H Theory and OJE as a leadership/managerial strategy and process simply subsumes our current program with a more comprehensive, substantive effort in an understandable framework. Teaching M-H Theory at the SNCO Academy, The Basic School, Amphibious Warfare School (intermediate level) and Command and Staff College is necessary. The leadership is then established for implementation on the jobs these leaders inherit. Finally, people and work issues are kept contemporary via the current 20 hour leadership package each Marine participates in annually. The institutionalization mechanism is in place now. Success in projects undertaken should fuel the fire of rapid expansion.

Finally what will be the modus operandi in project execution? The following is offered in response to that question and in summary of this paper:

1. Continue to educate down first to better facilitate changes originating at grass roots level. Many more "briefs of chiefs" must be conducted. A seminar of some 8-10 hours is in order covering theory, application, and executive level expectations.

2. Identify the two project organizations. This is easy. Volunteers abound. Several units have already attempted a piecemeal effort and are overwhelmed with positive response.

3. Select key-people for a 100-120 hour education process under Doctor Herzberg's tuteledge. These people must come from the project units--be flexible, knowledgeable in unit functions/

activities, each must display a capacity for managerial growth. Included in this group must also be a group of three or more people from HQ Marine Corps Manpower who will subsequently aid in the institutionalization process as well as provide guidance for new projects if any are desired. Doctor Herzberg will act only as a consultant on request by this time. The keyman is just that--the heart of the program. His selection is paramount to success.

4. Key-men can then conduct briefings for Executive Groups (if any) and Coordinating Groups usually 4-6 people.

INSERT CHART 8 ABOUT HERE

These are the middle supervisors; the commanders and/or staff who approve and support the efforts of the implementing groups.

5. Implementing groups where the work is done. These 6 to 8 people are the key supervisors--lieutenants, platoon sergeants or squad leaders. These are the people who know and represent the jobs to be enriched. Key-men train these people in detail.

This group then:

a. Blocks out the job and flow charts them.

b. Identify bad job characteristics.

c. Greenlight changes.

d. Redlight changes (good job characteristics installed).

e. Develop implementing plans for approval.

f. Work with Coordinating Group on a milestone chart for the implementing plan.

g. Determine at least interim indicies for both measurement and feedback.

6. Execute the plan:

   a. Measure results.

   b. Provide feedback.

   c. Refine and modify indicies.

By utilizing the above approaches and measures we should be able to track and evaluate the success and progress of our OJE effort at all levels.

REFERENCES

1. Herzberg, Fredericks - The Managerial Choice, To Be Efficient and to Be Human, Down Jones-Irvin, Homewood, Illinois, 1976

2. Orthodox Job Enrichment Handbook, Air Force Logistics Command, Ogden Utah, 1977

3. Has the Research Challenge to Motivation-Hygiene Theory Been Conclusive? An Analysis of Critical Studies, Gregagliunas and Wiemer, Human Relations, Volume 27 Number 9 pp 839-871

4. Herzberg, Frederick, Work and the Nature of Man, New York World Publishing Company, 1966

5. Ford, Robert N., Motivation Through The Work Itself, New York: American Management Association, Incorporated, 1969.

6. Herzberg, Frederick, Motivation-Hygiene Profiles, Organizational Dynamics, Fall, 1974.

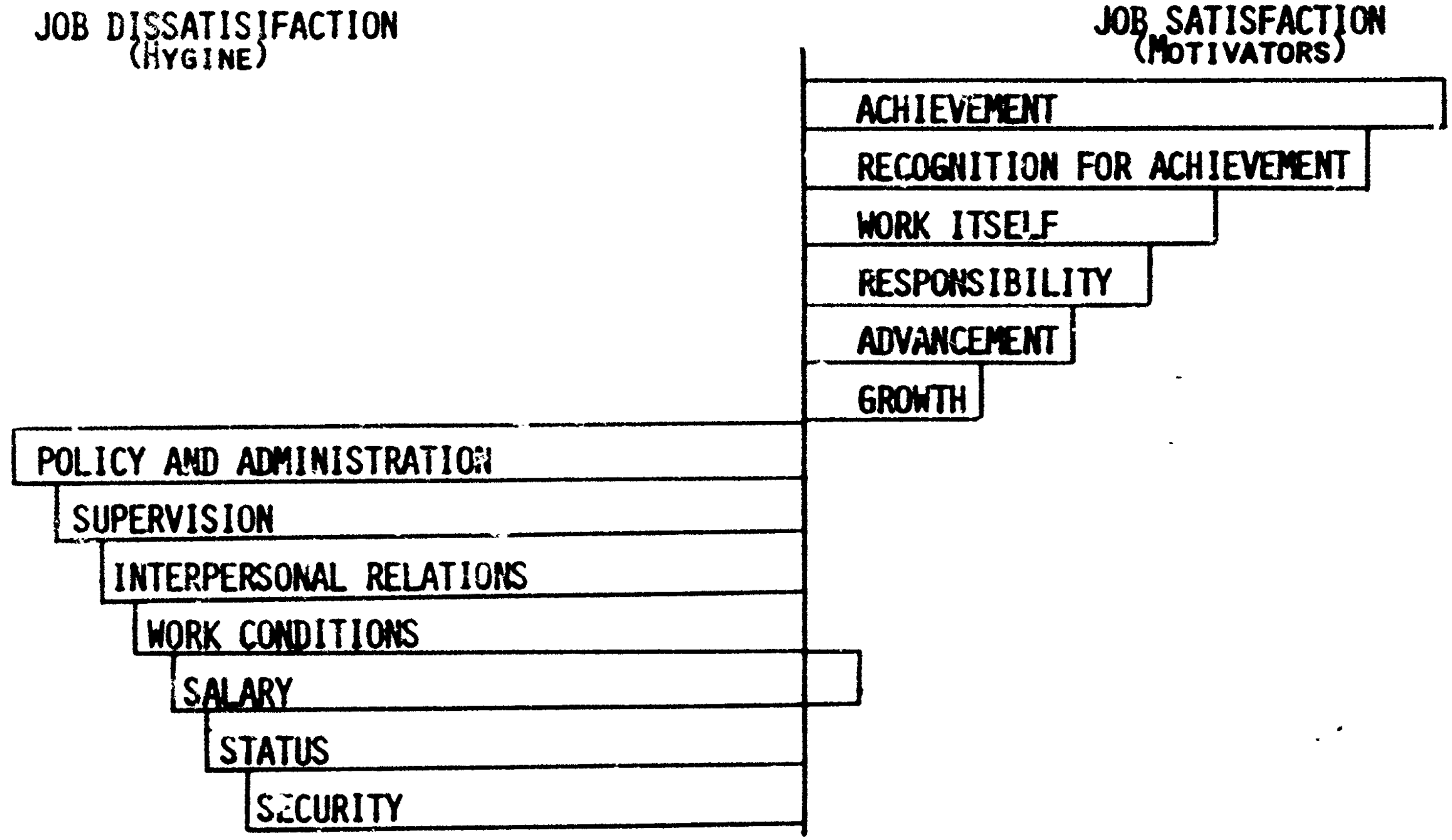
CHART 1

SALIENT DIFFERENCES

OLD DRAFTEE (PRE KOREA)	TODAY'S VOLUNTEER
1. OLDER	1. YOUNGER
2. HIGH SCHOOL AND SOME COLLEGE	2. HIGH SCHOOL (USUALLY)
3. ENTERED SERVICE RELUNCTANTLY AND THEREBY NOT SUBJECT TO PROFOUND DISILLUSIONMENT AFTER SERVICE ENTRY; ACCEPTED MILITARY ON ITS OWN TERMS.	3. ENTERS MILITARY AS AN ALTERNATIVE TO LIMITED OPTIONS IN CIVILIAN LIFE; REGARDS MILITARY IN INSTRUMENTAL TERMS OF "WHAT CAN IT DO FOR ME?" E.G., SKILL TRAINING, EDUCATION; SUBJECT TO POSTENTRY DISILLUSIONMENT IF EXPECTATIONS ARE NOT MET.
4. HIGHLY VALUE CLEAN RECORD WITH GOOD CONDUCT DISCHARGE.	4. IF DISILLUSIONED, WANTS OUT REGARDLESS OF TYPE OF DISCHARGE (THOUGH MAY LATER REGRET LACK OF GOOD DISCHARGE).
5. WILLING TO PUT UP WITH PETTY HARASSMENT.	5. MORE CONCERNED WITH SELF-DIGNITY; QUICKER TO TAKE EXCEPTION TO HARASSMENT.
6. REGARDED OVERSEAS ASSIGNMENT AS POTENTIALLY ENRICHING.	6. REGARDS OVERSEAS ASSIGNMENT AS AN IMPOSITION.

CHART 2

MOTIVATION-HYGIENE THEORY



NOTE: (1) ALL HYGIENE FACTORS ARE ARRANGED IN ORDER OF __FREQUENCY__ OF OCCURANCE NOT IMPORTANCE.
(2) MOTIVATORS ARE ALSO IN ORDER OF FREQUENCY AND INVERSE ORDER OF IMPORTANCE

CHART 3

DYNAMICS

HYGIENE	FACTOR	MOTIVATOR
1. DISSATISFACTION WHEN AT SOME LOW LEVEL. NO JOB DISSATISFACTION AT THE HIGH LEVEL.	EFFECTS	1. PERSONAL SATISFACTION THROUGH PSYCHOLOGICAL GROWTH WHEN PRESENT.
2. AVOIDANCE	TYPE OF BEHAVIOR	2. APPROACH
3. ENVIRONMENT	SOURCE	3. TASKS OR JOB CONTENT
4. INFINITE	NO OF SOURCES	4. FINITE OR LIMITED
5. LESS THAN	TYPE OF FEELING	5. MORE THAN
6. SHORT TERM	DURATION	6. LONG TERM
7. CYCLICAL	PSYCHOLOGICAL FUNCTION	7. ADDITIVE
8. ESCALATING	ZERO POINT	8. NONESCALATING
9. NONE-PAIN BUILT INTO BIOLOGY	OVERALL ANSWERS	9. MORE

CHART 4

CONCEPTUAL FORMULA FOR MOTIVATION

$$\text{MOTIVATION} = F\left(\frac{A}{P} : \frac{O}{A} : R\right)$$

WHERE A IMPLIES ABILITY

P IMPLIES POTENTIAL

O IMPLIES OPPORTUNITY

R IMPLIES REINFORCEMENT

CHART 5

INGREDIENTS OF BAD JOBS

1. LOW OR FEW SKILLS REQUIRED.
2. PERFECT PROCEDURES--FOLLOW DETAILED PROCEDURES USUALLY DEVELOPED AT A VERY LOW LEVEL/SKILL DENOMINATOR.
3. PERFECT TRAINING TO REINFORCE PERFECT PROCEDURES; USUALLY INDOCTRINATION VICE LEARNING.
4. FRAGMENTED TASKS.
5. HIGH LEVEL OF AUTOMATION IN THE ENVIRONMENT; RECOGNIZES THE HUMAN BEING AS A WEAK LINK.
6. COMPUTERIZED INPUT IS HIGH; THUS MINIMIZING HUMAN THINKING.
7. JOB MEASURED IN DETAIL; WHEREBY PEOPLE WORK FOR THE MEASUREMENT; INDEX TO MOVES/MOTION.
8. NARROW JOB CLASSIFICATION EXEMPLIFIED BY THE "THOU SHALT NOT" SYNDROME.
9. OVERLAPPING SUPERVISION.
10. EFFORTS ARE SCORED AGAINST PAR; INDICATES A SCORE AGAINST MEDIOCRACY.
11. SHORT-CYCLE TIME; TENDS TO MESMERIZE PEOPLE.

CHART 6

INGREDIENTS FO A GOOD JOB

A GOOD JOB SHOULD PROVIDE:

1. DIRECT FEEDBACK--QUICK, DIRECT AND NONEVALUATIVE.
2. A CLIENT RELATIONSHIP THAT MUST BE SERVED WELL; CAN BE PERSONAL, GEOGRAPHIC, ALPHABETICAL CUSTOMER OR ANY LOGICAL GROUPING.
3. NEW LEARNING AS A PART OF HIS JOB; IN ADDITION TO ANY TRAINING PROGRAM.
4. OPPORTUNITY TO SCHEDULE HIS OWN WORK; AVOID FORCED COOPERATION.
5. UNIQUE EXPERTISE; THIS HELPS PROVIDE SOME PSYCHOLOGICAL PRIVACY; HAVE OTHERS REFER TO HIM FOR THIS UNIQUE TALENT.
6. CONTROL OVER OWN RESOURCES; PROMOTES CONSCIOUSNESS OF ASSETS; GIVE HIM A BUDGET--SOMETHING TO WORK WITH, WITHIN, TOWARD.
7. DIRECT COMMUNICATION WITH THE CLIENTS/CUSTOMER; AVOID THE "CHECKER" SYNDROME EVIDENCED IN THE CLOAK OF SUPERVISION; ENCOURAGE QUICK, ACCURATE, RESPONSIBLE DECISION-MAKING BY THE PERSON "IN THE ACTION".
8. PERSONAL ACCOUNTABILITY; HOLD THE PERSON WHO DOES THE WORK ACCOUNTABLE FOR HIS WORK AND DO NOT HOLD HIM ACCOUNTABLE FOR DOING DUMB WORK; EACH PERSON SHOULD HAVE A RESPONSIBLE JOB AND BE HELD ACCOUNTABLE FOR IT; SCRUPULOUSLY AVOID EVALUATING PEOPLE ON THEIR PERSONALITY TRAITS WHICH IS A COMMON ERROR.

CHART 7

ENRICHED JOB

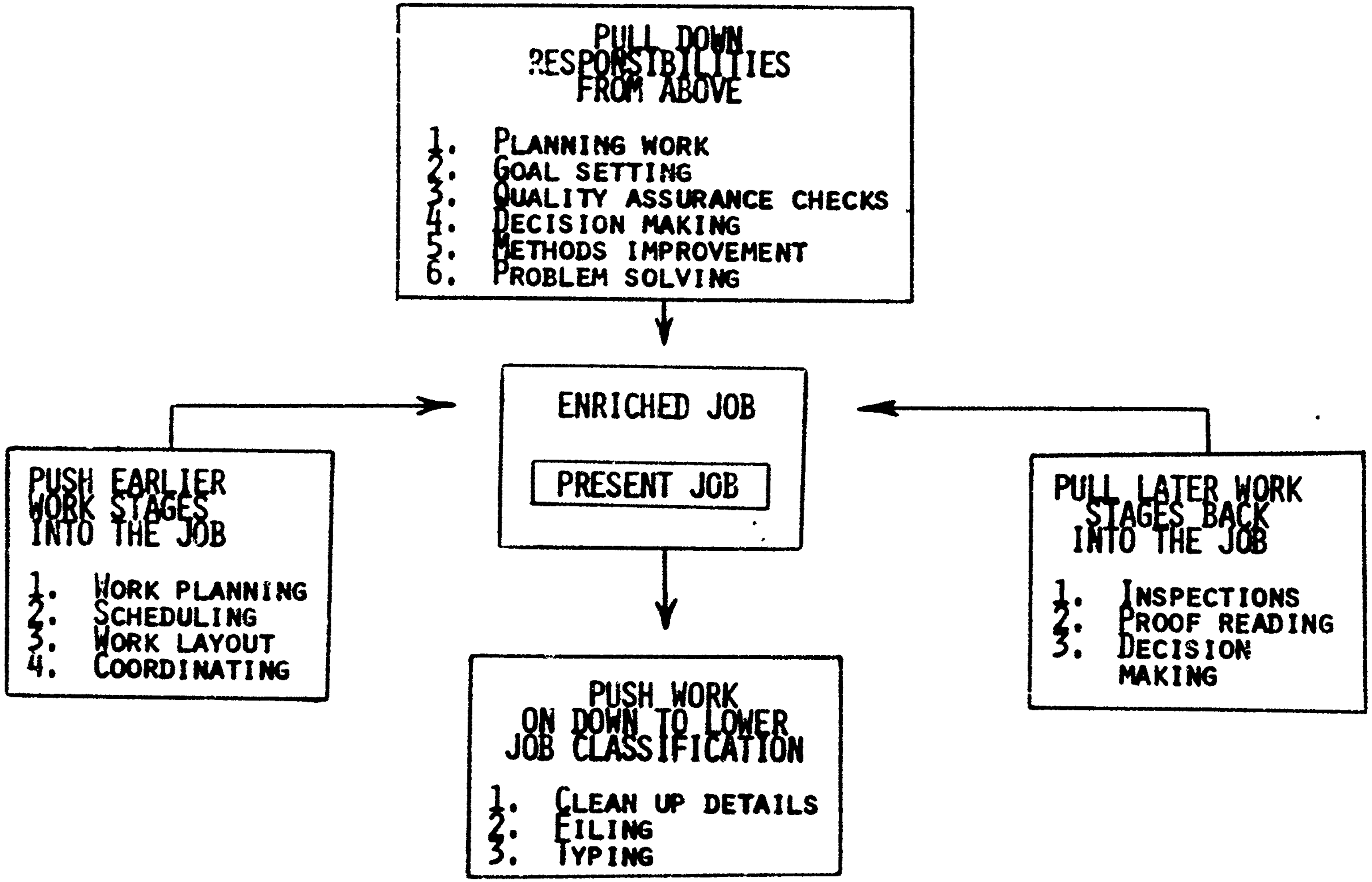
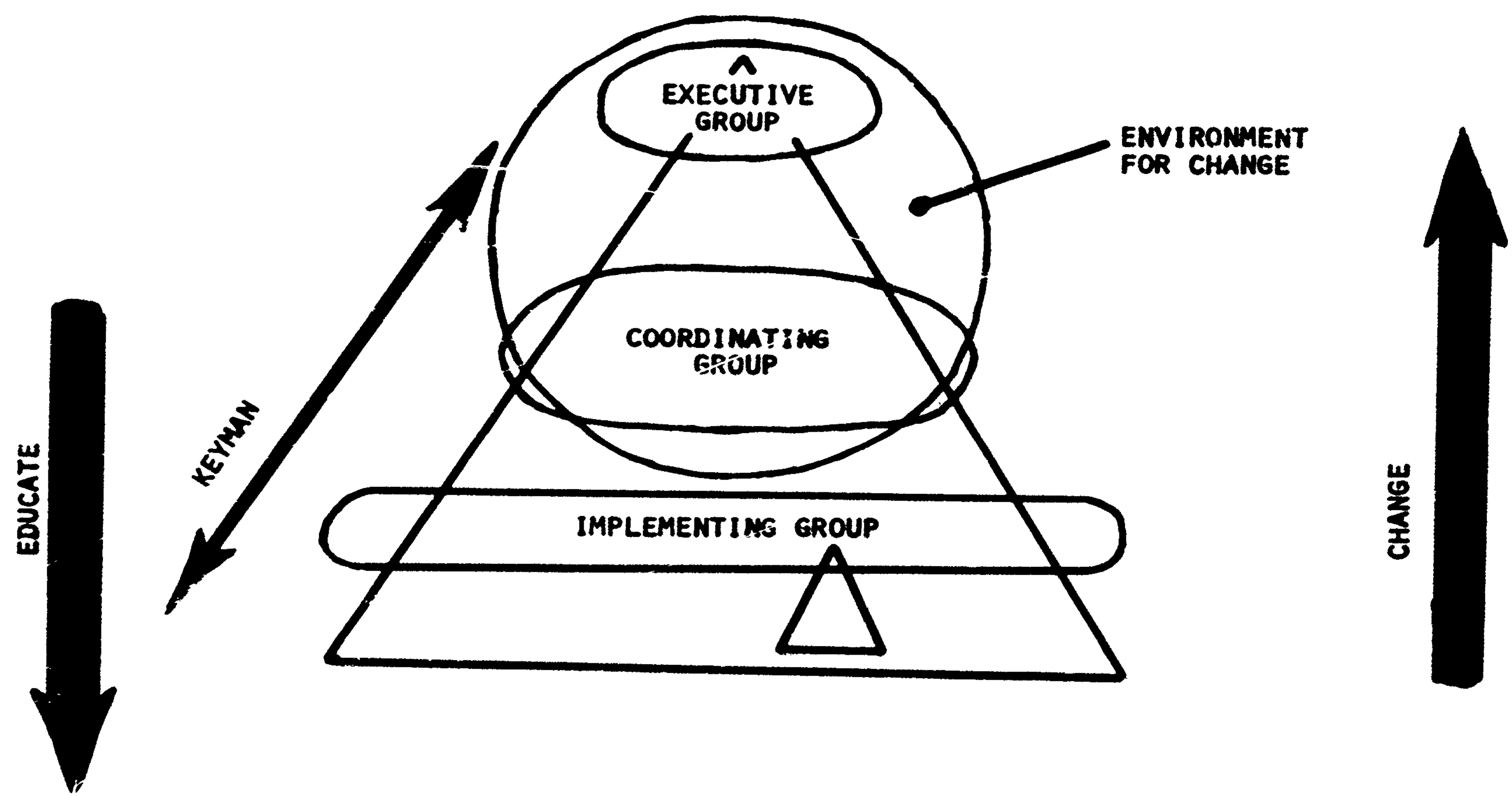


CHART 8

ORGANIZATION FOR ORTHODOX JOB ENRICHMENT



OVERVIEW OF THE US ARMY JOB SATISFACTION AND RETENTION PROJECT

MAJ DENNIS J. RUPP

JOB SATISFACTION AND RETENTION

MY PRESENTATION IS AN OVERVIEW OF THE ARMY'S JOB AND CAREER SATISFACTION PROJECT. MY INTENT IS TO PROVIDE A CONCEPTUAL AND FACTUAL FRAMEWORK SO YOU CAN BETTER UNDERSTAND AND EVALUATE OUR PROJECT AND PRESENTATIONS THIS MORNING.

I WILL PRESENT THE OVERVIEW IN THREE PARTS: THE CONTEXT IN WHICH THE PROJECT BEGAN, THE INTENDED USES OF THE DATA, AND THE COMPONENTS OF THE PROJECT.

I WOULD PREFER THAT THIS BE MORE OF AN INFORMAL DISCUSSION THAN A HIGHLY FORMAL, LECTURE-LIKE PRESENTATION. SO IF YOU HAVE ANY QUESTIONS OR COMMENTS, PLEASE TOSS THEM OUT AS THEY ARISE.

CONTEXT

IN JUNE 1976, MILPERCEN BEGAN EXPANDING THE JOB SATISFACTION PORTION OF THE OCCUPATIONAL SURVEY PROGRAM. THE PURPOSE OF THE EXPANSION WAS TO BETTER UNDERSTAND HOW SATISFACTION WITH ONE'S ARMY JOB AND MILITARY LIFE AFFECTS THE DECISION TO STAY OR LEAVE THE SERVICE. WE ESPECIALLY HAVE BEEN INTERESTED IN THE RELATIONSHIP BETWEEN JOB AND CAREER SATISFACTION AND FIRST-TERM REENLISTMENTS.

THE MILPERCEN PROJECT IS PART OF THE ARMY'S INTENSIFIED EFFORT TO GAIN ADDITIONAL INSIGHTS INTO RETENTION,

JOB SATISFACTION, AND THE ALL-VOLUNTEER ARMY. THE ULTIMATE OBJECTIVE IS TO IMPROVE THE ARMY'S ABILITY TO RECRUIT AND RETAIN A SUFFICIENT NUMBER OF QUALITY SOLDIERS.

AS TUTTLE AND HAZEL HAVE POINTED OUT, MOST OF THE RESEARCH AND APPLICATIONS OF JOB SATISFACTION HAVE OCCURRED IN INDUSTRY (TUTTLE AND HAZEL, 1974). HOWEVER, WITHIN THE LAST TEN YEARS THE MILITARY SERVICES HAVE BEGUN TO APPLY THE RESEARCH FINDINGS FROM THE PRIVATE SECTOR AND TO SPONSOR THEIR OWN RESEARCH IN THIS AREA. MUCH OF MILPERCEN'S WORK BUILDS ON RELATED RESEARCH BY THE ARMY RESEARCH INSTITUTE (ARI) AND THE AIR FORCE'S HUMAN RESOURCES LABORATORY (AFHRL). THE JOB AND CAREER SATISFACTION PROJECT MOST CLOSELY RESEMBLES THE AIR FORCE'S APPROACH (ALLEY AND GOULD, 1975). DR. LARRY GOLDMAN WILL EXPLAIN MORE ABOUT THIS IN HIS PRESENTATION.

ALTHOUGH OCCUPATIONAL ANALYSIS BEGAN SYSTEMATICALLY ARMY-WIDE IN 1968, THE JOB SATISFACTION PORTION WAS NOT ADDED TO THE OVERALL SURVEY PROGRAM UNTIL FALL 1974.

TWENTY ITEMS WERE USED TO OPERATIONALLY DEFINE AND EMPIRICALLY MEASURE SATISFACTION WITH ONE'S ARMY JOB AND WITH MILITARY LIFE. THE OPERATIONAL DEFINITIONS OF JOB AND CAREER SATISFACTION WERE BASED ON THE HYGIENE AND MOTIVATOR FACTORS THAT FREDERICK HERZBERG IDENTIFIED IN HIS RESEARCH ON JOB SATISFACTION (HERZBERG, MAUSNER, AND

SNYDERMAN, 1959). THIS VUGRAPH (SEE FIGURE A BELOW) LISTS HERZBERG'S MOTIVATOR AND HYGIENE FACTORS. THE ASTERISKED VARIABLES ARE THOSE FOR WHICH SPECIFIC OCCUPATIONAL SURVEY ITEMS HAD BEEN WRITTEN.

FIGURE A

HERZBERG'S MOTIVATORS	HERZBERG'S HYGIENE FACTORS
ACHIEVEMENT	COMPANY POLICY & ADMINISTRATION
* RECOGNITION	* SUPERVISION
WORK ITSELF	* RELATIONSHIP WITH SUPERVISOR
* RESPONSIBILITY	* WORK CONDITIONS
* ADVANCEMENT	* SALARY
* GROWTH	* RELATIONSHIP WITH PEERS
	* PERSONAL LIFE
	RELATIONSHIP WITH SUBORDINATES
	* STATUS
	SECURITY

ALTHOUGH THE HERZBERG-BASED ITEMS WERE ADEQUATE FOR OCCUPATIONAL ANALYSIS, WE DECIDED TO DISCARD THIS APPROACH IN FAVOR OF ONE THAT WAS DESIGNED TO HELP PREDICT CAREER DECISIONS, WORK ATTITUDES, AND DUTY PERFORMANCE (ALLEY AND GOULD, 1975).

INTENDED USES

THE TWO [illegible] USES OF THE HERZBERG-BASED JOB [illegible]

FACTION DATA WERE TO IDENTIFY THE RELATIVE DEGREE OF SATISFACTION/DISSATISFACTION AMONG DIFFERENT OCCUPATIONAL GROUPS AND TO AMPLIFY ON RELATED DATA COLLECTED IN OTHER PARTS OF THE OCCUPATIONAL SURVEY QUESTIONNAIRE. THE UNDERLYING PURPOSE OF THESE EFFORTS WAS TO PROVIDE OCCUPATIONAL INFORMATION TO THE ARMY'S TRAINERS AND OCCUPATIONAL STRUCTURES PEOPLE.

IN REASSESSING THE INTENDED USES OF THE JOB AND CAREER SATISFACTION PROJECT, TWO CRITERIA WERE FOLLOWED:

- EXPANDED USES WOULD BE BASED ON THE CURRENT AND FUTURE NEEDS OF KEY ARMY DECISION-MAKING AGENCIES.
- THE PROJECT WOULD BE LINKED TO OTHER RELATED RESEARCH AND STUDY EFFORTS IN THE ARMY, OTHER SERVICES, AND INDUSTRY.

CONSEQUENTLY, CONSIDERABLE TIME AND EFFORT WAS DEVOTED TO DETERMINING THE NATURE AND SCOPE OF OTHER ON-GOING WORK AND IDENTIFYING THE NEEDS OF KEY AGENCIES.

THE OUTCOME OF THIS REASSESSMENT WAS AN EXPANDED LIST OF INTENDED USES. THIS VUGRAPH (SEE FIGURE B BELOW) SHOWS THE EXPANDED USES.

FIGURE B

EXAMINE RELATIONSHIP BETWEEN JOB SATISFACTION AND:

- RETENTION
- MORALE
- OCCUPATIONAL MISMATCH
- EFFECTIVE USE OF TRAINED ASSETS
- SELECTED STUDIES (E.G. WOMEN IN THE ARMY)

JOB AND CAREER SATISFACTION SURVEYS

OUR PROJECT CONSISTS OF FOUR PROJECT ELEMENTS: THREE ARMY-WIDE ATTITUDE SURVEYS AND ONE OCCUPATIONAL SURVEY. I SHALL EXPLAIN EACH OF THESE BRIEFLY.

IN AUGUST '76 A SAMPLE SURVEY FOCUSED ON FIRST-TERM SOLDIERS IN GRADES E-3 AND E-4. APPROXIMATELY 3,700 FIRST-TERMERS RESPONDED TO 38 QUESTIONS ABOUT JOB SATISFACTION AND REENLISTMENT INTENT. THE RESULTS WERE PUBLISHED AS A SURVEY REPORT IN MAY 1977. THE REPORT IS ENTITLED "JOB SATISFACTION AND REENLISTMENT INTENT FOR FIRST-TERM PERSONNEL: INITIAL FINDINGS."

IN FEBRUARY OF THIS YEAR WE RANDOMLY SAMPLED APPROXIMATELY 4,000 FIRST-TERM AND CAREER ENLISTED MEN AND WOMEN. THE SURVEY CONSISTED OF 80 ITEMS ABOUT JOB SATISFACTION AND CAREER DECISIONS. THE FEBRUARY SURVEY WAS AN ABBREVIATED VERSION OF THE AIR FORCE'S OCCUPATIONAL ATTITUDE INVENTORY. IT WAS MODIFIED TO ACCOUNT FOR THE DIFFERENCES BETWEEN THE ARMY AND AIR FORCE. THE RESULTS OF THIS SURVEY WILL BE PUBLISHED NEXT MONTH. MR. DARRELL WORSTINE IN HIS PRESENTATION WILL HIGHLIGHT SOME OF THE MAJOR FINDINGS OF THE FEBRUARY '77 SURVEY.

IN MARCH OF THIS YEAR WE SENT OCCUPATIONAL SURVEY QUESTIONNAIRES TO 1,725 RECRUITERS AND CAREER COUNSELORS (MOS 00E). THIS SURVEY WILL PROVIDE INSIGHT INTO THEIR

PERCEPTIONS ABOUT THE CAREER MOTIVATION OF ENLISTED SOLDIERS. THE RESULTS OF THE SURVEY WILL BE PUBLISHED EARLY NEXT YEAR.

AT THE END OF THIS MONTH, WE WILL RANDOMLY SURVEY APPROXIMATELY 30,000 FIRST-TERM AND CAREER ENLISTED MEN AND WOMEN. THESE SOLDIERS WILL BE ADMINISTERED A COMPREHENSIVE JOB AND CAREER SATISFACTION QUESTIONNAIRE (APPROXIMATELY 300 ITEMS). THE QUESTIONNAIRE REPRESENTS THE CULMINATION OF MORE THAN ONE YEAR OF DEVELOPMENTAL WORK. THE RESULTS OF THE OCTOBER 1977 SURVEY WILL BE PUBLISHED BY JUNE OR JULY 1978.

THE AUGUST 1976 ARMY-WIDE SURVEY AND THE APRIL 1977 PILOT TEST

DR LAWRENCE A. GOLDMAN

I THE AUGUST 1976 ARMY-WIDE SURVEY

A. INTRODUCTION

IN AUGUST 1976, AN 80 ITEM ARMY QUARTERLY SAMPLE SURVEY WAS DISTRIBUTED TO A RANDOM SAMPLE OF PERSONNEL ARMY-WIDE. SINCE ITEMS INCLUDED IN THIS SURVEY WERE FINALIZED PRIOR TO INITIATION OF THE JOB SATISFACTION/REENLISTMENT INTENT PROJECT, THIS QUESTIONNAIRE WAS NOT DESIGNED TO BE A COMPREHENSIVE INSTRUMENT FOR MEASURING SPECIFIC FACTORS INFLUENCING SOLDIERS' ATTITUDES TOWARDS THESE TWO CRITERION MEASURES. IN PARTICULAR, COVERAGE OF FACTORS WITH THE POTENTIAL OF INFLUENCING REENLISTMENT WAS INCOMPLETE. JOB SATISFACTION WAS ADDRESSED PRIMARILY BY 17 INDEPENDENT FACTORS DEVELOPED BY THE MILITARY OCCUPATIONAL DEVELOPMENT DIVISION, FOR WHICH DATA HAVE BEEN COLLECTED SYSTEMATICALLY FROM ENLISTED PERSONNEL THROUGH MILITARY OCCUPATIONAL DATA BANK QUESTIONNAIRES SINCE SEPTEMBER 1974. TO EXTEND THE DOMAIN OF MEASUREMENT INTO OTHER AREAS BELIEVED TO INFLUENCE JOB SATISFACTION AND POSSIBLY REENLISTMENT, AN ADDITIONAL 21 ITEMS INCLUDED IN THIS QUARTERLY SAMPLE SURVEY (NOT BE PERSONNEL AFFILIATED WITH THIS PROJECT) WERE ALSO CONSIDERED IN THE ANALYSIS. THE MAJOR DEFECTS IN THE COVERAGE OF REENLISTMENT RELATED FACTORS IN THE AUGUST 1976 QUESTIONNAIRE HAVE BEEN REDUCED TO A CONSIDERABLE EXTENT IN THE FEBRUARY 1977 ARMY-WIDE SURVEY TO BE DISCUSSED BY MR. WORSTINE. AS MAJOR HUPP INDICATED IN HIS INTRODUCTORY REMARKS, ANALYSIS OF THE AUGUST 1976 SURVEY

WAS CONDUCTED ON APPROXIMATELY 3,700 PERSONNEL IN PAYGRADES E-3 AND E-4 WHO WERE IN THEIR INITIAL TERM OF ENLISTMENT.

B. SIGNIFICANT FINDINGS AND CONCLUSIONS

1. INTERESTING WORK WAS IDENTIFIED AS THE BEST PREDICTOR OF BOTH REENLISTMENT INTENT AND JOB SATISFACTION, AT LEAST FOR FIRST TERM PERSONNEL. THIS FINDING WAS NOTED FOR E-3'S AS WELL AS E-4'S, MALES AND FEMALES, NON-HIGH SCHOOL GRADUATES AND HIGH SCHOOL GRADUATES, WHITES AND BLACKS, AND SINGLE AND MARRIED PERSONNEL. FROM THIS IT COULD BE INFERRED THAT THE EXTENT TO WHICH SOLDIERS PERCEIVE THEIR WORK TO BE INTERESTING STRONGLY INFLUENCES OVERALL SATISFACTION WITH THEIR JOB AND THEIR INTENTION TO REENLIST. THIS IS INDEED SIGNIFICANT SINCE IT IMPLIES THAT A NON-MONETARY FACTOR PLAYS A GREATER ROLE IN JOB SATISFACTION AND REENLISTMENT INTENT THAN MILITARY PAY, ALLOWANCES AND/OR BENEFITS, A BELIEF COMMONLY SHARED BY MANY CIVILIANS AS WELL AS INDIVIDUALS WITHIN THE ARMY. WITH THE CONCOMITANT RESPONSIBILITY OF THE ARMY TO REDUCE OVERALL PERSONNEL-RELATED COSTS WHILE INCREASING THE RETENTION RATE OF ELIGIBLE PERSONNEL, ESPECIALLY UNDER THE ALL VOLUNTEER FORCE, ATTEMPTING TO MAKE JOBS MORE APPEALING MAY ALLEVIATE THESE TWIN CONCERNS.

2. WORK IMPORTANCE, WORK CHALLENGE, AND WORKING ASSOCIATION WITH ONE'S SUPERVISORS WERE RELATIVELY CONSISTENT PREDICTORS OF JOB SATISFACTION IN TERMS OF GRADE, SEX, EDUCATIONAL LEVEL, RACE, AND MARITAL STATUS. IN OTHER WORDS, A SOLDIER PERCEIVES A JOB TO BE SATISFYING IF HE/SHE BELIEVES THAT IT IS TRULY SUBSTANTIAL WHILE

ESSENTIALLY FREE FROM "HARRASSMENT" BY HIS/HER SUPERVISOR.

3. SOLDIERS WHO BELIEVED THEY WERE GIVEN ACCURATE INFORMATION BY AN ARMY RECRUITER HAD HIGHER REENLISTMENT INTENTIONS AND GREATER JOB SATISFACTION THAN THOSE WHO DIDN'T. SOLDIERS WHO FEEL THEY WERE GIVEN ACCURATE INFORMATION BY AN ARMY RECRUITER PROBABLY ENTER THE ARMY WITH A MUCH CLEARER IDEA OF WHAT TO EXPECT FROM ARMY LIFE. THIS DOES NOT IMPLY THAT ARMY RECRUITERS EITHER TRULY REPRESENTED OR MISREPRESENTED THE FACTS ABOUT ARMY LIFE. WHAT THIS REPRESENTS IS THE EXTENT TO WHICH THE EXPECTATION OF THE INDIVIDUAL CORRESPONDED TO THE INFORMATION IMPARTED TO HIM/HER BY THE ARMY RECRUITER. THIS, IN TURN, MAY BE A PRIMARY REASON FOR THE INCREASED LIKELIHOOD THAT HE/SHE HAS A SIGNIFICANTLY HIGHER PROPENSITY TO REENLIST AND JOB SATISFACTION.

II THE APRIL 1977 PILOT TEST

A. INTRODUCTION

IN THE EARLY PLANNING PHASE OF THE OVERALL PROJECT, SEVERAL WEEKS WERE SPENT DISCUSSING THE NATURE AND SCOPE OF AN ARMY JOB AND CAREER SATISFACTION MODEL. THE ORIGINAL INTENT WAS TO DEVELOP A MODEL AND TEST IT. HOWEVER, BECAUSE OF TIME AND MANPOWER CONSTRAINTS, IT WAS DECIDED TO CAPITALIZE ON THE EXTENSIVE LITERATURE REVIEW AND LONG-RANGE RESEARCH CONDUCTED BY THE US AIR FORCE HUMAN RESOURCES LABORATORY CONCERNING JOB/CAREER SATISFACTION. NOT ONLY DID THE AIR FORCE'S EFFORTS PARALLEL MILPERCEN'S, BUT THEIR RESEARCH PROGRAM WAS CLOSELY LINKED TO AND BASED UPON AN OCCUPATIONAL ANALYSIS PROGRAM AND ONE OF

THE MAIN PURPOSES OF THEIR RESEARCH WAS TO STUDY THE RELATIONSHIP BETWEEN JOB SATISFACTION AND RETENTION (REENLISTMENT). SINCE THE AIR FORCE CONCLUDED THERE WERE NO ADEQUATE JOB SATISFACTION MEASUREMENT INSTRUMENTS FOR USE IN THE MILITARY ENVIRONMENT, AN OCCUPATIONAL ATTITUDE INVENTORY (OAI) WAS DEVELOPED. IN THE INITIAL DEVELOPMENT OF THE OAI, 36 POTENTIAL SATISFACTION DIMENSIONS (ALSO REFERRED TO AS HYPOTHESIZED FACTORS) WERE IDENTIFIED BY AIR FORCE BEHAVIORAL SCIENTISTS FAMILIAR WITH THE MILITARY WORK ENVIRONMENT. ITEMS WERE WRITTEN FOR EACH DIMENSION, RESULTING IN A FINAL POOL OF 348 ITEMS OF APPROXIMATELY 10 ITEMS PER DIMENSION. TO VALIDATE THESE HYPOTHESIZED FACTORS, A RANDOM SAMPLE OF 3,000 FIRST TERM AIRMEN WAS ANALYZED.

IN THE INITIAL DEVELOPMENT OF AN ARMY PILOT TEST QUESTIONNAIRE, ALL OF THE AIR FORCE'S HYPOTHESIZED FACTORS WERE USED EXCEPT "PERSONAL GROWTH AND DEVELOPMENT", "INDEPENDENCE" AND "UNCLASSIFIED". IN ADDITION TO ADOPTING THE BASIC AIR FORCE'S DIMENSIONS, FOUR FACTORS WERE HYPOTHESIZED FOR THE PILOT TEST "FAMILY", "INDIVIDUAL", "DISCRIMINATION", AND "ARMY UNIQUE". THESE FOUR ADDITIONAL FACTORS, BASED ON THE ADULT DEVELOPMENT RESEARCH OF DANIEL LEVINSON AND ROGER GOULD, ARE BELIEVED TO BE IMPORTANT INFLUENCES ON A PERSON'S MOTIVATION AND BEHAVIOR. THE FACTORS HYPOTHESIZED BY THE ARMY AND THE NUMBER OF ITEMS ASSOCIATED WITH EACH FACTOR ARE SHOWN IN THE TABLE.

\- - - - - - - - - - - - -

INSERT THE TABLE ABOUT HERE

\- - - - - - - - - - - - -

TABLE

AIR FORCE		ARMY	
FACTOR DESCRIPTOR	NUMBER OF ITEMS	FACTOR DESCRIPTOR	NUMBER OF ITEMS
ACHIEVEMENT	7	ACHIEVEMENT	2
ACTIVITY	8	ACTIVITY	4
AIR FORCE AND UNIT POLICIES AND PRACTICES	18	ARMY AND UNIT POLICIES AND PRACTICES	17
ASSIGNMENT LOCALITY	17	ASSIGNMENT LOCALITY	16
AUTHORITY	4	AUTHORITY	3
CO-WORKERS	9	CO-WORKERS	12
CREATIVITY	10	CREATIVITY	5
IMPORTANCE	8	IMPORTANCE	2
INTEREST	9	INTEREST	4
KNOWLEDGE OF RESULTS	7	KNOWLEDGE OF RESULTS	3
PERSONAL GROWTH AND DEVELOPMENT	9	——	
JOB DESIGN	10	JOB DESIGN	3
OPTIONAL SOCIAL CONTACT REQUIRED SOCIAL CONTACT	7 10	SOCIAL CONTACT	11
PAY AND BENEFITS	12	PAY AND BENEFITS	8
PHYSICAL WORK ENVIRONMENT	13	PHYSICAL WORK ENVIRONMENT	9
PROMOTION OPPORTUNITY	8	PROMOTION OPPORTUNITY	4
RECOGNITION	9	RECOGNITION	4
RESPONSIBILITY	10	RESPONSIBILITY	4
INDEPENDENCE	9	——	
VALUE OF EXPERIENCE	8	VALUE OF MILITARY EXPERIENCE	3

AIR FORCE		ARMY	
FACTOR DESCRIPTOR	NUMBER OF ITEMS	FACTOR DESCRIPTOR	NUMBER OF ITEMS
PHYSICAL SAFETY	6	PHYSICAL SAFETY	3
ECONOMIC SECURITY	4	ECONOMIC SECURITY	2
SERVICE TO OTHERS	8	SERVICE TO OTHERS	1
SOCIAL STATUS	11	SOCIAL STATUS	5
SUFFICIENCY OF TRAINING	12	SUFFICIENCY OF TRAINING	10
SUPERVISION RECEIVED - HUMAN RELATIONS	15	HUMAN SUPERVISION	16
SUPERVISION RECEIVED - TECHNICAL	9	TECHNICAL SUPERVISION	5
PERFORMANCE EVALUATION	8	PERFORMANCE EVALUATION	3
JOB CHANGE	7	JOB CHANGE	4
TOOLS, EQUIPMENT AND SUPPLIES	8	TOOLS, EQUIPMENT, AND SUPPLIES	7
UTILIZATION	8	UTILIZATION	3
VARIETY	9	VARIETY	4
WORK SCHEDULE	15	WORK SCHEDULE	6
SUPERVISORY DUTIES	18	SUPERVISORY DUTIES	10
UNCLASSIFIED	8	——	
——		INDIVIDUAL	14
——		ARMY UNIQUE	10
——		DISCRIMINATION	9
——		FAMILY	18
TOTAL	348	TOTAL	244

In general, those items were selected for inclusion in the pilot test questionnaire which loaded highest under each of the Air Force's 35 empirically derived factors using principal components analysis. Also, most of the original items utilized by the Air Force were modified to reflect the differences in Army and Air Force environments and terminology. Efforts were also made to enhance item clarity.

Of the original 348 items used by the Air Force, 208 items were retained, either in their original wording or in modified form. A total of 26 items were added to the hypothesized Air Force factors while another 90 were included for the four new factors used in the pilot test questionnaire. The combined total of 324 items was then reduced to 225 based on the following criteria: (a) Redundancy of items; (b) Reducing the number of items in each of the hypothesized factors "Individual", "Human Supervision", and "Family"; and (c) Constraints imposed by the length of the answer sheet. All items were measured on a seven point scale, ranging from EXTREMELY DISSATISFIED to EXTREMELY SATISFIED.

The pilot test was administered by three teams of two individuals each from MILPERCEN to an availability sample of approximately 1,600 personnel at six US installations. In addition, roughly 600 individuals were personally interviewed. These interviews provided insights into the content validity of the questionnaire and allowed soldiers the opportunity to comment on the wording, length, sensitivity and coverage.

B. The Pilot Test as Foundation for the Army-Wide Survey

1. Effect of Order of Presentation of the 225 Job/Career Satisfaction Variables. Since the soldiers responding to the pilot were asked to respond to 225 items (74 items in each of three sections and three in another section), it was hypothesized that the mean value of each of these variables might vary according to where it appeared in the questionnaire. To test this hypothesis, these three sections were counterbalanced (i.e., the order of the three sections was varied at the six installations.) One way analysis of variance of the six mean values obtained for each variable indicated statistically significant differences for only 12 or 5.4 percent of the 222 separate tests. Thus, there were no substantial differences in terms of the effect of the order of presentation.

2. Reduction of the 225 Job/Career Variables to 124 for the Army-Wide Survey. The 225 items were reduced to a final total of 124 through application of the following three procedures: (a) factor analysis; (b) stepwise multiple regression analysis and discriminant function analysis; and (c) subjective review, primarily based on the previously obtained results.

The 225 variables were factor analyzed with a primary objective of identifying those variables whose loadings were "significantly high" (i.e., those whose correlations with any of the rotated factors was .40 or higher). The 225 items considered as independent variables were then examined through stepwise multiple regression and discriminant function

ANALYSIS. FOR BOTH THESE ANALYSES, THE DEPENDENT VARIABLES USED TO GAIN INTO THE MULTI-FACETED ASPECTS OF REENLISTMENT BEHAVIOR AND SATISFACTION COMPRISED THE FOLLOWING: (A) REENLISTMENT PLANS; (B) JOB SATISFACTION; (C) ARMY SATISFACTION; (D) DESCRIPTION OF UNIT MORALE; AND (E) DESCRIPTION OF ONE'S JOB. A TOTAL OF 121 ITEMS LOADED SIGNIFICANTLY ON AT LEAST ONE FACTOR AND WERE SIGNIFICANT PREDICTORS AND/OR DISCRIMINATORS OF AT LEAST ONE OF THE FIVE CRITERION MEASURES UTILIZED. FIFTEEN ITEMS WERE ADDED SO THERE WOULD BE AT LEAST ONE ITEM FOR EACH OF THE 35 HYPOTHESIZED FACTORS. THE 136 ITEMS WERE THEN SUBJECTIVELY REVIEWED TO ELIMINATE DUPLICATION WITHIN THE SAME HYPOTHESIZED FACTOR. ALSO, SEVERAL ITEMS WERE ELIMINATED WHICH WERE JUDGED TO BE OF LITTLE PRACTICAL VALUE IN TERMS OF JOB/ARMY CAREER SATISFACTION, MORALE, AND RETENTION (E.G., "YOUR OPINION OF THE ARMY COMPARED TO THE AIR FORCE"). STARTING FROM THE 136 ITEMS BASED ON OBJECTIVE ANALYSIS, SUBSEQUENT AND SUBJECTIVE REVIEW REDUCED THIS NUMBER TO THE 124 ITEMS CONSTITUTING SECTION B OF THE COMPREHENSIVE ARMY-WIDE JOB AND CAREER SATISFACTION SURVEY.

Implementation of a Model Adaptive Testing System at an Armed Forces Entrance and Examination Station

Malcolm James Ree
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

In a world of increasing technical complexity and diminishing resources, it is the task of the military recruiting agencies to obtain the most highly qualified candidates for technical training. Traditionally, paper-and-pencil multiple aptitude test batteries have been administered to applicants of a wide range of abilities. These tests have been peaked to be most discriminating over a relatively narrow range because limited time precluded the administration of enough items to gain maximal test information over a broad range of an ability. However, selection and classification decisions must be made which require discriminations at the 80th percentile. At this level, only limited information is available from a peaked test.

Adaptive testing, particularly computer-driven adaptive testing, promises to enable the gathering of test information (Lord and Novick, 1968, eq 20.2.7) at all levels of ability with equal precision and to increase the predictive validity of our military accession testing. Furthermore, adaptive testing promises to reduce the time required to gain ability estimates for applicants and thus possibly reduce overall costs by making accession a one day process.

The model adaptive testing system was implemented in an Armed Forces Entrance and Examination Station (AFEES) in order to study its feasibility

for use in a military selection setting. At the AFEES, the testing system must be operated by individuals without any special training in computer hardware or software. The system must perform when needed; it must be operational for the entire workday, and also accommodate applicants for military service from very low ability to very high ability. It must not intimidate or frighten the applicants or the test administrators. Finally, it must provide valid and reliable measurement.

Prior to the implementation of an adaptive testing system, many decisions must be made, both technical and administrative. The technical questions include: who are the subjects, what ability areas are to be tested, what items and item statistics are available, which scoring method will be used, which item selection technique will be used, what media for question presentation will be used, and how pictorial items will be presented.

There are also many administrative questions. How can the operation be simplified so that low ability, careless, or inattentive examinees do not cause an abnormal ending of the program? What impact will the demonstration have on day to day AFEES operations?

The San Antonio, Texas, AFEES was chosen as the test site because it was close to the development center at the Air Force Human Resources Laboratory (AFHRL) which afforded considerable opportunity for monitoring the progress of the adaptive testing system.

The subjects for this demonstration were applicants for military enlistment, and their abilities covered a very broad range of aptitudes. They were tested in three aptitude areas which comprise the Armed Forces

Qualification Test (AFQT): Word Knowledge (WK), Arithmetic Reasoning (AR), and Space Perception (SP). The AFQT is used for initial qualification for military service. Other aptitude areas are usually measured only if an acceptable score on the AFQT is achieved. The subjects were tested while awaiting the results of the AFQT.

The items used for this model adaptive testing system were culled from existing historic item files at the AFHRL. Only item difficulty (P) and item discrimination (Phi) indices were available. Items were selected to represent a generally rectangular distribution of difficulties from about .2 to about .8 with the highest available discrimination index at each difficulty level. These items were then assembled into booklets for administration to Air Force basic recruits in order to estimate latent trait parameters a, b, and c (Lord & Novick, 1968) for later phases of this demonstration. Initially, the classical item indices were transformed via approximations and used to calculate the latent trait parameters useful for the project. Although these estimates would vary somewhat from the more exact estimates obtained from the new response data, they did permit a reasonable starting point from which to demonstrate the feasibility of adaptive testing for applicants for military service. As soon as a satisfactory sample has been collected and the parameters estimated, the approximated parameters will be replaced by the new, more exact parameters.

A medium size computer, IBM 360/65, was available for the demonstration in a time-shared mode. The APL programming language (Gilman & Rose, 1970) was selected because it is interactive and has extremely powerful

operators. In addition, experience has shown that APL leads to fast development. It is also fast in execution and is particularly suited for handling vectors and matrices.

A combination of Bayesian item scoring/ability estimation (Owen, 1969) and selection of items by maximum information (Lord & Novick, 1968, eq 20.4.1) was selected for ease of programming and low computer core utilization. These are two criteria for termination item administration: reduction of the posterior variance of the ability estimate to a low value (> .0625) and/or the subjects having taken 20 items. This procedure is also advantageous because it does not require a structured item pool as would a stratified adaptive test thus making implementation of the testing system easier.

A modified Tektronix model 4006-1 Cathode Ray Tube (CRT) terminal was used for the demonstration. A viewing hood to reduce glare and a keyboard cover to prohibit pushing inappropriate keys were fabricated. This terminal supported the Tektronix Graphics Package, APLgraph 2, and was run at 1200 BAUD in half duplex mode.

In order to insure proper operation of the system, operating instructions and operating safeguards were built in. The examinee is taught how to use and respond to the terminal before any questions are presented. All solicitations for input are for characters ('1,' '2,' '3,' '4,' '5'), as opposed to numbers, and are checked to determine the presence of alphabetic (ABCD, etc.) or special characters ($!&, etc.). If an out-of-range response, an alphabetic character, or a special character is given, the instructions for responding are repeated. Then the screen is cleared and finally the

question is repeated. Proper character input is then converted to its equivalent numerical form and processed.

As expected, the characters for the questions of WK and AK are kept on an external randomly accessible file and read in as needed. Screen control characters are stored with the literal characters which makes for simplicity of operation. The last array of the file, roughly equivalent to the last record of a FORTRAN file, contains a 4 by N (N = the number of items in the file) matrix of the item parameters and answer keys. This matrix is read in and manipulated prior to the presentation of all questions.

Producing the pictorial displays for the SP items presented a unique problem in storage and drawing. One proposed storage method was to use back screen projection from a random access slide projector. This was discounted because it allowed only about 100 items to be stored, added a mechanical component to maintain, and required photographic slides of each item. Similarly, the idea of writing a specific mathematical function to generate each individual figure was dismissed because it required extensive programming for each new item. Finally, a method of display was developed using the sophisticated graphics capability of APL which only requires placing a drawing of the SP item on a "digitizing tablet" and touching various points on the drawing. These are then transformed into a vector for each item, and the graphics package draws the many lines represented by the vector at a very rapid rate (see Figure 1). At 1200 BAUD, the figures almost flash onto the screen as the vector is read in from its place on the file. This technique could be extended to any item requiring drawings, such as mechanical principles or block counting.

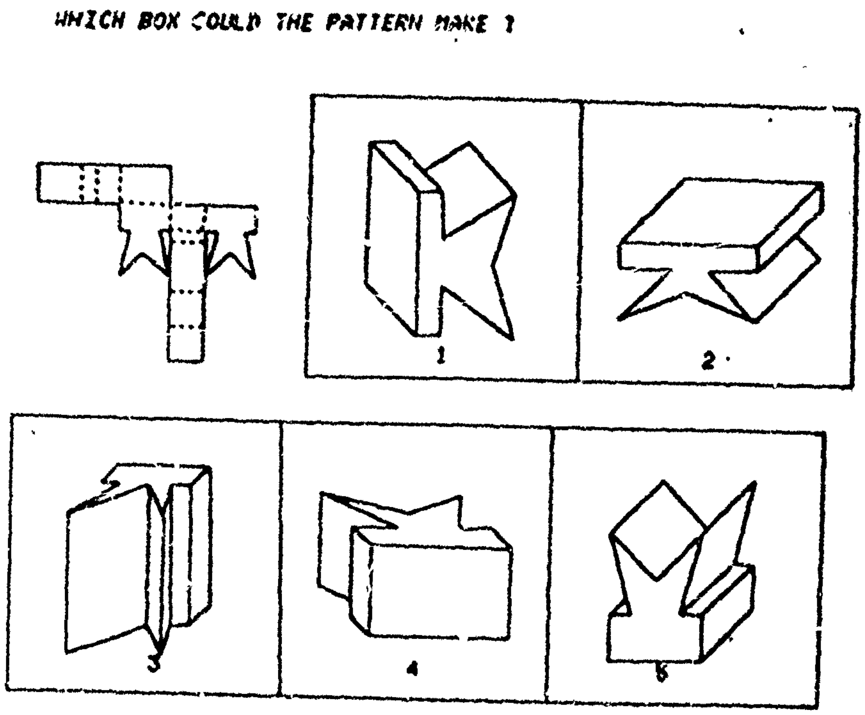


Figure 1. Direct copy of a typical Space Perception item from screen of CRT.

All technology and software developed to draw the Space Perception items are general enough to be used in other testing or educational applications. The perspective and three dimensional effect are very good, and motion for rotating or shifting the figures can be added. Rotating figures, demonstration of mechanical principle, or moving lever arms may lead to new item types not amenable to static paper-and-pencil tests. Computer driven graphics may allow us to measure new and important ability areas.

An operator's manual, algorithmicly written, was produced for the AFEES personnel. It contains complete instructions for initial daily starting and stopping of the testing system. It also gives instructions for starting the program if the terminal is already running. The manual offers names and telephone numbers of people to contact in the event of trouble. The programs have been "locked" to the AFEES staff, and they have been advised not to try to edit the programs. Back-up copies of both the programs and the files are stored on line and require only a command from the proper user to reinstate damaged programs or to update programs as they are refined.

Data grade telephone lines, a special telephone number for the AFEES use only, and a special sign-on code were provided to reduce competition for telephone ports in the time-shared environment. The "Special Testing Room" at the AFEES was used to house the terminal. This is a 10' x 12' windowless room containing several student chairs with arms, one side chair, and a 3' x 2' table for the terminal. The terminal and the telephone connector need little space and can be operated in any room with 117 volts AC and a telephone.

The feasibility of adaptive testing will be investigated in this demonstration by assessing two important factors. First, did the system run with little trouble and attention? This will be assessed from interviews with the AFEES staff and from daily logs of the system's operation. Secondly, was adaptive testing as valid as paper-and-pencil testing? The validity of the adaptive testing system will be assessed by comparing the subjects' adaptive scores and the subjects' AFQT subtest scores. Analysis of these data will help in making future decisions about adaptive testing.

Following this demonstration there will be questions to answer before any large scale implementation can be undertaken. Some of these questions are psychometric, some logistic, and some economic. As yet, no testing configuration, local or nationwide, has been developed, nor have system costs for implementing, operating, and supporting adaptive testing been established. Basic conceptual questions dealing with such diverse topics as testing models, back-up systems, operating policies, and central versus dispersed processing remain unanswered.

It is conceivable that certain other decisions will facilitate broad scale implementation of adaptive testing. For example, the AFEES in Baltimore, Maryland, already has computer-automated management and paper handling on an in-house mini-computer. The addition of adaptive testing might require little additional hardware, and, in quantity, this additional hardware might be inexpensive enough to merit its use. Furthermore, adaptive testing could add to test security because neither test booklets nor answer key are

distributed, and no one can have knowledge beforehand as to which questions will be administered to a subject.

In the future, the actual costs and benefits of adaptive testing will be known. This will permit realistic decision making for its use. This knowledge will allow adaptive testing to move from the fad of the 1970's to the operational tool of the 1980's and beyond.

References

Gilman, L., & Rose, A. APL 360 An Interactive Approach. New York, John Wiley & Sons, Inc., 1970.

Lord, F., & Novick, M. Statistical Theories of Mental Test Scores. Reading MA: Addison-Wesley Publishing Co., 1968.

Owen, R. A Bayesian Approach to Tailored Testing. Princeton NJ: Educational Testing Service, Research Bulletin RB 69-92, 1969.

AN ADAPTIVE TEST OF ARITHMETIC REASONING

James R. McBride

In recent years there has been growing interest among test theoreticians and practitioners in adaptive, or tailored, ability testing as an alternative to group-administered conventional tests. The reasons for this interest have been many, but a key reason is the psychometric efficiency of tailored tests: by tailoring the choice of test items to the individual test-taker, a short, well-designed adaptive test can match the measurement precision of a much longer conventional test. In theory, it is possible for an adaptive test to equal a conventional test's reliability (and validity) in less than half its length.

The theoretical advantages of adaptive tests are not realized without some cost, however. These costs can be expressed in item quality and quantity. Urry (1970) demonstrated that the reliability/validity advantages of adaptive tests depended on the availability of unusually highly discriminating test items -- items with discriminating power equivalent to item-trait biserial correlations exceeding .62. Lord's (1970) theoretical analyses of adaptive tests focused on branching procedures such as the stair-step (Lord, 1974) or pyramidal (Larkin & Weiss, 1975) procedure, and the Robbins-Munro procedure. The stair-step method required an item pool containing $\frac{1}{2}(\underline{k}(\underline{k}+1))$ items in order to administer an individualized $\underline{k}$-item test to each examinee; thus a 15-item pyramidal test required a 120-item pool of highly discriminating test items. The superior Robbins-Munro tailoring procedure required $2^{k}-1$ items in the pool for an individual test length of $\underline{k}$ items; a 15-item Robbins-Munro test would require 32767 items in the pool!

Other adaptive strategies made less exorbitant but still stringent item pool demands. Jensema (1977) recommended an item pool size exceeding 100 items in order to implement Owen's Bayesian adaptive strategy. Urry (1974) screened about 900 operational test items in order to assemble a 200-item pool to measure verbal ability. In short, the number and quality of items seemingly required to implement an adaptive testing strategy raised serious questions about the feasibility of using adaptive tests in settings where item resources are limited. The purpose of this paper is to describe an attempt to construct an useful adaptive test from a limited number of available test items. The attempt was a theoretical one, but as you will see below it was motivated by a practical problem, and was based on analysis of the psychometric properties of real test items. You will also see below that the attempt was successful, a fact which should have important implications for future practice.

BACKGROUND

The mental testing portion of the military enlistment screening process consumes about three hours of each examinee's processing time. That time is used to administer the Armed Services Vocational Aptitude Battery (ASVAB), which consists of twelve cognitive subtests and an interest inventory. If the time required to administer the cognitive tests could be reduced substantially, some of the available three hours could be used profitably in other ways: to collect biographical data, for example, or to assess reading skills. However, reduction of that testing time would require either eliminating some subtests, or shortening some or all of them; an alternative solution, if it were feasible, might be to use adaptive tests of the cognitive abilities. Part of the feasibility question hinges on the availability of sufficient numbers of highly discriminating test items having a wide distribution of difficulty. The purpose of the analyses reported below was to assess the feasibility of constructing short adaptive subtests of ASVAB, using available items, without detriment to the psychometric quality of the test scores.

The Arithmetic Reasoning (AR) subtest was taken as a case in point. AR is a 20-item subtest in the currently operational ASVAB Forms 6 and 7, is highly reliable for its length, and considered singly is one of the best subtests in terms of validity with external criteria. A target objective of the feasibility study was to determine whether an adaptive version of AR could be devised which would use available test items, have psychometric properties equal to an operational AR test, yet be only half as long. In order to specify the target objective rigorously, an information analysis (Birnbaum, 1968) of the AR subtest of ASVAB Form 6 was performed, using approximations of the parameters of each of the AR subtest's 20 items' characteristic curves. The resulting test information function is illustrated in Figure 1.

For those of you who are unfamiliar with this kind of analysis, let me refer you to Birnbaum (1968) for a detailed presentation, but state briefly here that the test information function is an index of the test's measurement precision as a function of location of the ability scale. The higher the local value of the test information function, the more useful is the test for discriminating among examinees in that region of the scale. The test information function is related to the test reliability (Samejima, 1977); unlike the reliability coefficient, however, the information function is invariant from group to group. One important property of it is its relationship to the conditional variance of errors of measurement. Asymptotically, the conditional error variance equals the inverse of the information function; hence (again, asymptotically) the conditional standard error of measurement equals the inverse of the square root of the information function. Figure 2 is a graph of the inverse of the square root of the values graphed in Figure 1, and may be interpreted as illustrating the measurement error characteristics

of the 20-item operational AR test. It and Figure 1 are standards against which any substitute for the 20-item test may be judged.

THE ADAPTIVE TEST

Constructing an adaptive AR test required identifying a source of test items to stock the item pool and choosing a rationale for adaptive item selection. The obsolete Forms 2 and 3 of ASVAB were chosen as the item source, resulting in a pool of 50 AR items; this is a small item pool relative to the sizes usually suggested for adaptive testing. The item characteristic curve parameters of each item were approximated from available item analysis data.

The adaptive item selection rationale, or "strategy" (Weiss, 1974) chosen was a two-stage variant of a multi-level test strategy similar to that described by Lord (1977). This was motivated in part by the small size of the item pool, and in part by the desire that the test be administrable in paper-and-pencil form as well as by computer.

The first-stage, or "routing" test, is a three-item branching test based on a seven-item subset of the 50-item pool. The remaining 43 items were used to construct several seven-item overlapping levels of a multi-level test. Each examinee answers questions at just one level; the choice of level is based on his performance on the routing test.

The Routing Test

Figure 3 is a schematic diagram of the routing test. Every examinee answers item 1, which was chosen as the item which is locally optimal by the least-squares criterion proposed by Owen (1969) for item selection in a sequential Bayesian tailored testing procedure.

Before the test begins, the only information available about an examinee is the population mean ability, 0.0. Using the Bayesian procedure given by Owen (1969), we can update that information after observing performance on item 1. This results in two possible ability estimates: .40 after a right answer, versus -.94 for a wrong answer; associated with each score is an appraisal of its standard error. The combination of the score and standard error data permit us to choose two locally optimum items (one for each score on item 1) using the least-squares criterion. Each examinee is directed to the appropriate one of those two items. The same ability estimation and item selection procedure is used after the examinee answers the second item, to route him to one of four optimum third items. After the third item is answered, the ability estimation procedure results in

just eight different estimates -- one ability estimate corresponding to each of the eight possible patterns of item scores. Thus, a unique ability estimate is implied by the pattern of the examinee's scores on the three items comprising his routing test; assignment of the examinee to a level of the multi-level second stage test is based on that implied ability estimate.

The routing test may be computer-administered, in which case it is identical in scoring and item selection to Owen's Bayesian sequential tailored testing method. It will retain these properties if paper-and-pencil administration is used, and will have several important advantages over previously proposed paper-and-pencil branching tests. The primary advantage is that the branching task is quite simple: from the first item to the second, from the second to the third, and from the third item to one level of the multi-level test (each level of which may be printed on a separate page in the test booklet). The second advantage is that the test is optimally scorable even if an examinee makes an error in branching; this is so because the test is item response theory based; each item in the pool has had its item characteristic curve parameters estimated in advance, so that the test may be computer-scored using maximum likelihood estimation, with all scores expressed in a common metric regardless of what particular set of items the individual examinee has answered.

The Multi-Level Test

After seven items from the 50-item pool had been reserved for the routing test, forty-three items remained from which to construct the different levels of the multi-level second-stage test. Each level required seven items, but some item overlap was considered desirable to minimize the seriousness of routing errors.

There could be as many as eight levels -- one for each ability estimate resulting from the routing test. Items were assigned to levels using the "cut-and-try" method; the effect of each trial was analyzed psychometrically. It was finally determined that due to the small item pool there was no benefit to using more than six levels.

Figure 4 shows the allocation of test items to the six level tests; each item's psychometric characteristics are indicated by its location in the two-dimensional plane formed by item characteristic curve difficulty (horizontal axis) and discrimination (vertical axis). The arrows below the horizontal axis indicate the scale values of each of the eight possible routing test scores. The cluster of seven items constituting the level test corresponding to each routing test score is indicated on the figure.

Information Analyses

The combination of eight routing tests, and six different levels of the second-stage test, results in an adaptive test which administers one of eight different combinations of ten items to each examinee. The test information function of each 10-item combination was computed separately, along with the conditional probability of that combination occurring for specified ability levels. The overall information function of the adaptive test was computed from that data, using the formula:

$$I(\theta) = \sum_{k=1}^{8} I_k(\theta)\, P(v_k|\theta),$$

where

$I(\theta)$ = the adaptive test information function value;

$I_k(\theta)$ = the information function value of combination $\underline{k}$ ($1 \leq k \leq 8$);

$P(v_k|\theta)$ = the conditional probability of response score pattern $\underline{k}$ on the routing test.

$I_k(\theta)$ is calculated directly from the item characteristic curve parameters of the 10 items in a given combination $\underline{k}$. $P(v_k|\theta)$ is calculated from the item characteristic curve parameters of the routing test leading to combination $\underline{k}$.

The resulting information function is depicted in Figure 5. The graph of the corresponding conditional standard error curve is in Figure 6, overlaid with the counterpart curve for the 20-item conventional test (repeated from Figure 2). Figure 6 can be interpreted in the following manner: The conventional AR test achieves its lowest levels of measurement error in the region of the ability scale between 100 and 120; on the Army Standard Score scale this is the range between the population mean and one standard deviation above the mean. If a conditional standard error value of 7.0 is taken as tolerable (i.e., a standard error of measurement corresponding to a reliability coefficient of .88), the conventional test has satisfactory measurement properties from about 90 (one-half standard deviation below the mean) to 130 (one and one-half S.D.'s above the mean). It is not satisfactory in the range from 70 to 80, which is perhaps the most crucial region of the scale for Army enlistment screening purposes.

The curve of measurement error for the adaptive test shows it to be satisfactory throughout the range from about 75 to 136 on the Standard Score scale. It is notably superior to the conventional test in the critical range from 70 to 80, which implies that it should be a better test for screening prospective enlistees than is the operational AR subtest of ASVAB Form 6. The adaptive test's measurement error characteristics are inferior to those of the conventional test between 100 and 120 on the score

scale; the differences are slight, however, and the involved region of the scale is not critical for most Army screening or classification purposes.

Another way to compare the 10-item adaptive test with the 20-item conventional one is by means of a relative efficiency analysis. The relative efficiency (RE) index is simply the ratio of one test's information function to that of the other test. Since the adaptive test is here being considered as an alternative to the conventional one, its relative efficiency should equal or exceed 1.0 throughout the range of interest on the ability scale. Figure 7 is a plot of the relative efficiency of the adaptive AR test, compared to the conventional one. It tells the same story as Figure 6: The adaptive test has better measurement properties than the operational AR test throughout the important range of the standard score scale, with the exception of the region from the mean to one standard deviation above the mean; where it is only slightly inferior to the conventional test.

CONCLUSIONS

The information, conditional standard error, and relative efficiency analyses above show that it is feasible to construct a 10-item adaptive test of Arithmetic Reasoning which has overall measurement properties as good as those of a conventional test twice as long. It should be just as feasible for the other non-speeded ASVAB subtests. By virtue of the relationship between the information function and the reliability coefficient pointed out by Samejima (1977), the information analyses reported above imply that, for the military mobilization population, the short adaptive test should have reliability (and hence validity) at least equal to that of the longer, operational test.

The results reported above have important implications for the feasibility of implementing adaptive testing in the military selection setting. One implication is that successful adaptive counterparts of today's operational screening and classification tests can be implemented, using relatively small item pools. Since the 50-item pool used for this study was composed solely of items from now obsolete operational test forms, a second implication is that the quality and distribution of test items constituting our operational tests are quite adequate for adaptive testing purposes. A third implication, this one following from the simplicity of the adaptive testing procedure which was used, is that in principle it is feasible to develop adaptive tests which can be administered in paper-and-pencil form, without excessively complex instructions, and without the need for scoring intermediate stages of the test.

References

Birnbaum, A. - Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. Statistical Theories of Menital Test Scores. Boston: Addison-Wesley, 1968.

Jensema, C. - The influence of item bank characteristics on Bayesian testing. Applied Psychological Measurement, 1977.

Larkin, K.C. & Weiss, D.J. - An empirical investigation of computer-administered pyramidal ability testing. RR 74-3. Psychometric Methods Program, Department of Psychology, University of Minnesota, July 1974. (AD 783553).

Lord, F.M. - Some test theory for tailored tests. In W.H. Holtzman (ed.) Computer-assisted instruction testing and guidance. New York: Holt, Rinehart and Winston, 1970.

Lord, F.M. - Individualized testing and item characteristic curve theory. In Krantz, Atkinson, Luce & Suppes (Eds.). Contemporary developments in mathematical psychology. San Francisco, CA: W.H. Freeman, 1974.

Lord, F.M. - Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Owen, R. - A Bayesian sequential approach to tailored testing. RB 93-69. Princeton, N.J.: Educational Testing Service, 1969.

Samejima, F. - A use of the information function in tailored testing. Applied Psychological Measurement, 1977.

Urry, V.W. - A Monte Carlo investigation of logistic mental test models. Doctoral dissertation, Purdue University, 1970. (University Microfilms No. 71-9475).

Weiss, D.J. - Strategies of Adaptive Ability Measurement. RR 74-5. Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974. (AD 004279).

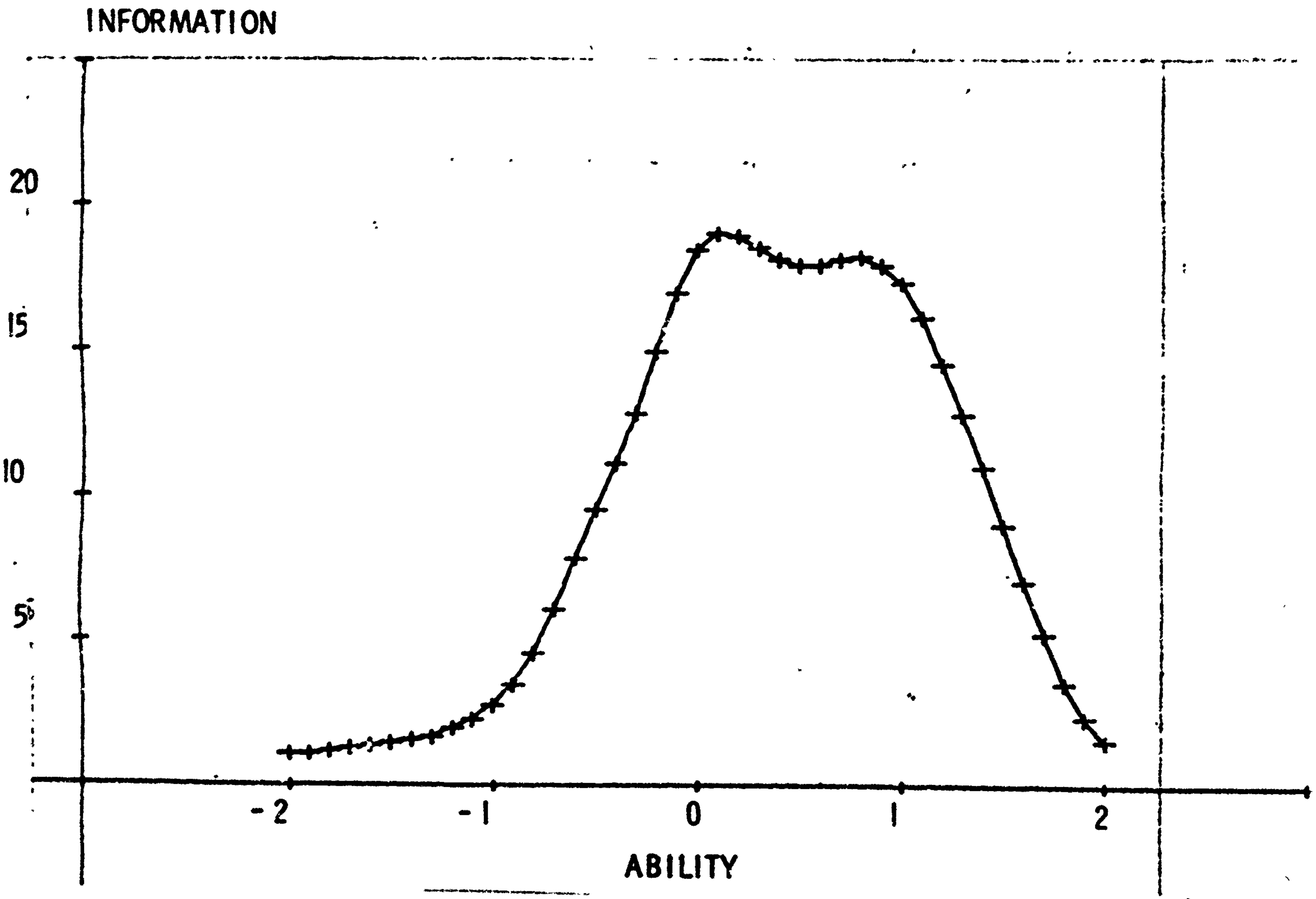


Figure 1. The test information function estimated for subtest AR, ASVAB Form 6. The ability scale is in standard deviation units.

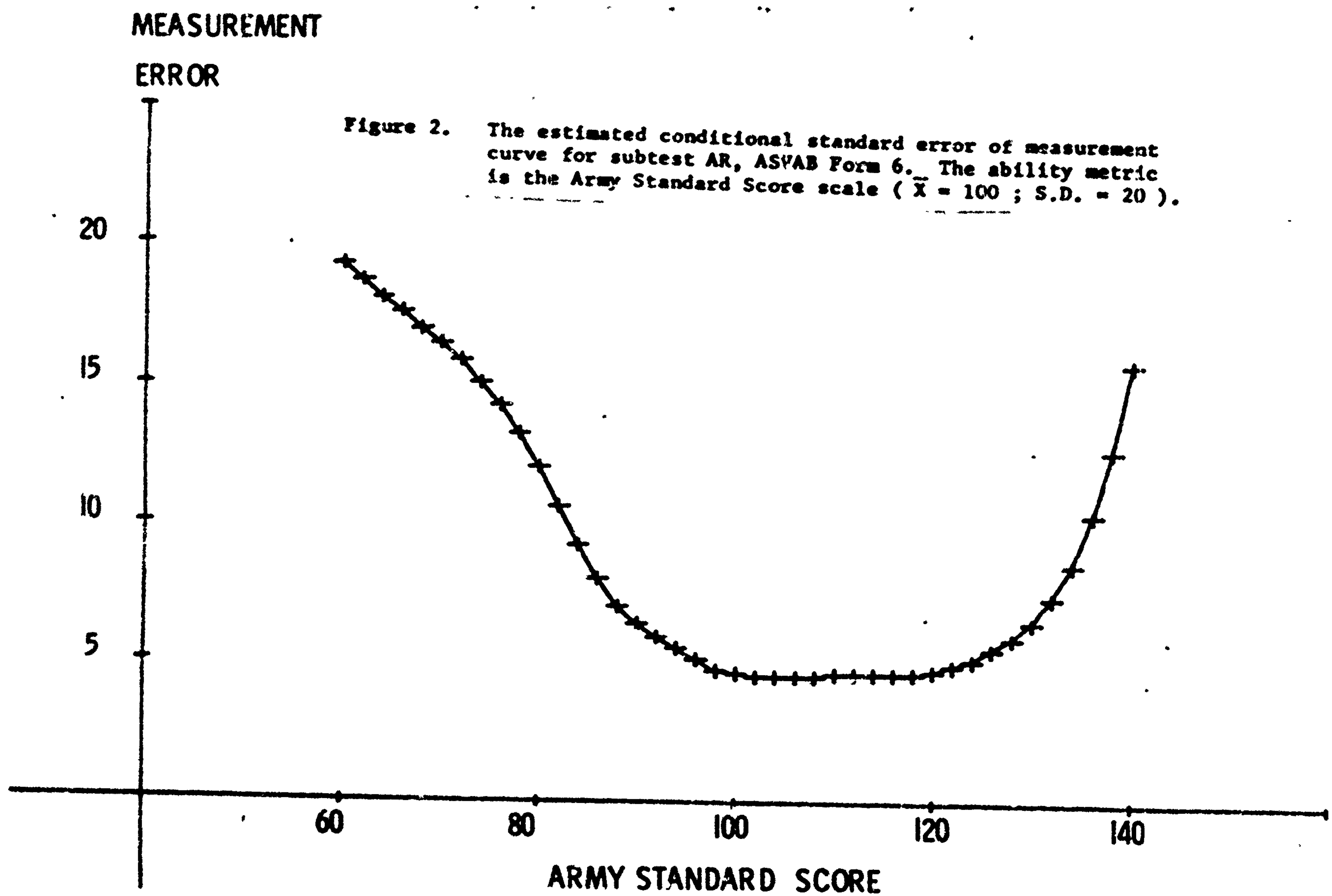


Figure 2. The estimated conditional standard error of measurement curve for subtest AR, ASVAB Form 6. The ability metric is the Army Standard Score scale ($\bar{X}$ = 100 ; S.D. = 20).

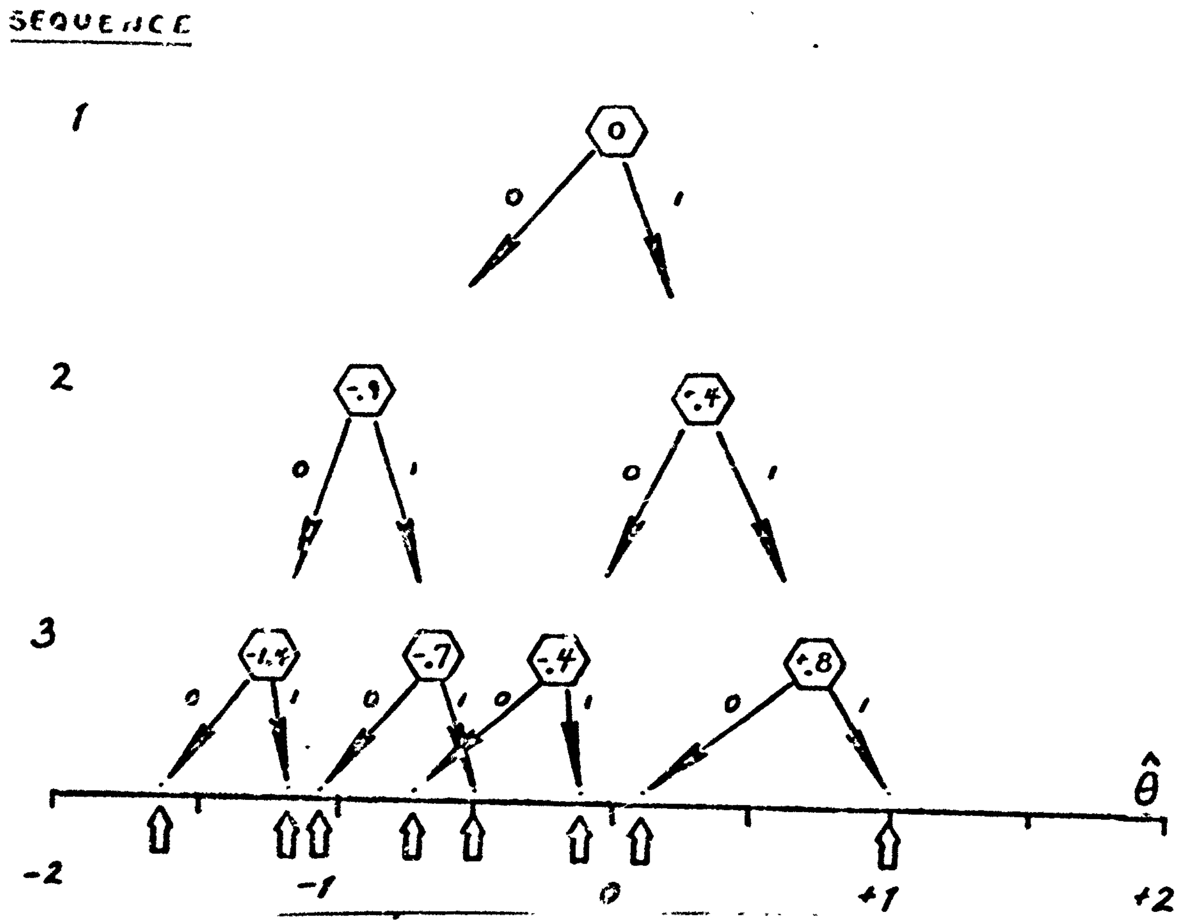


Figure 3. A schematic of the branching rule for the 3-item Bayesian adaptive routing test.

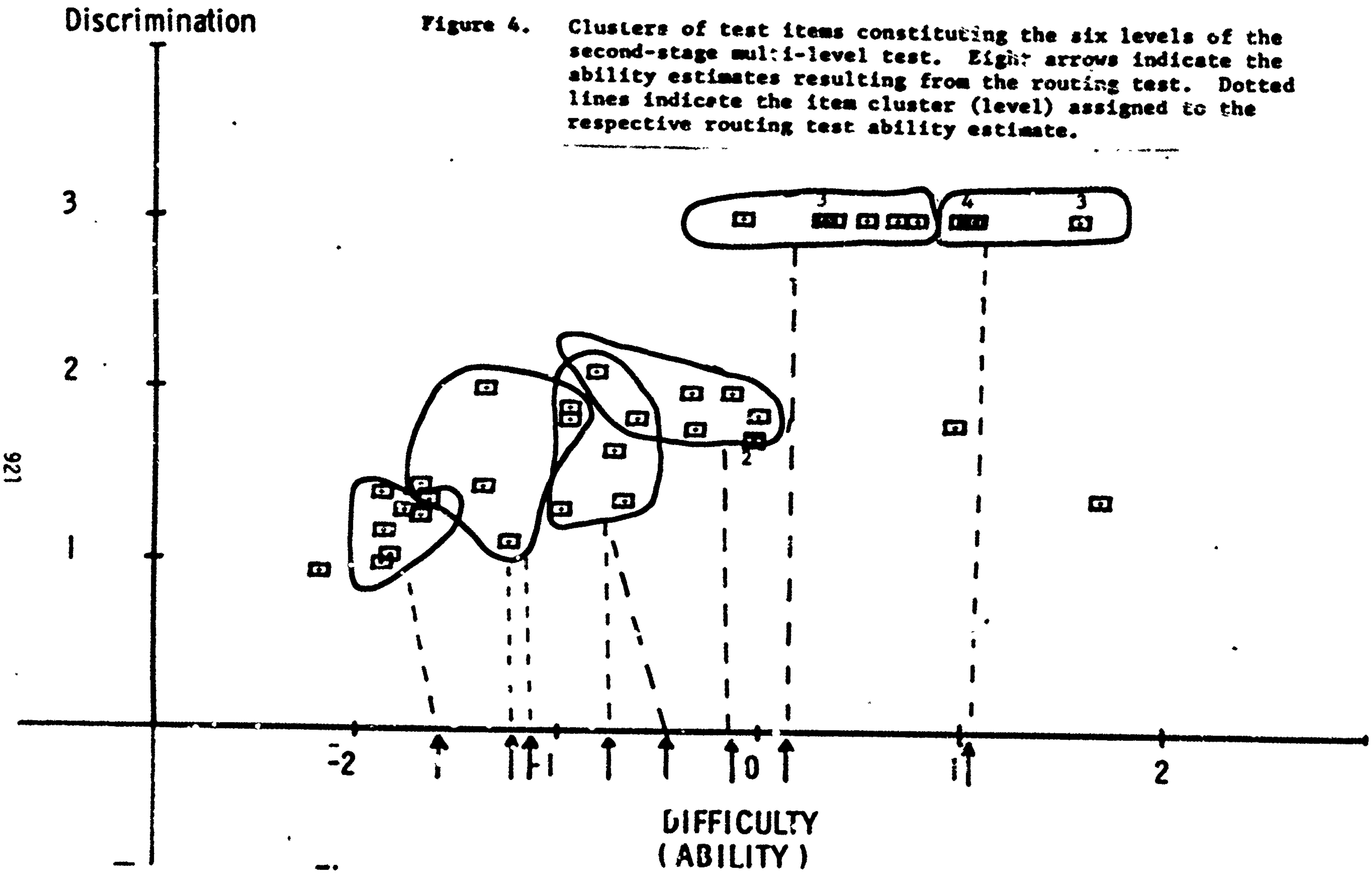
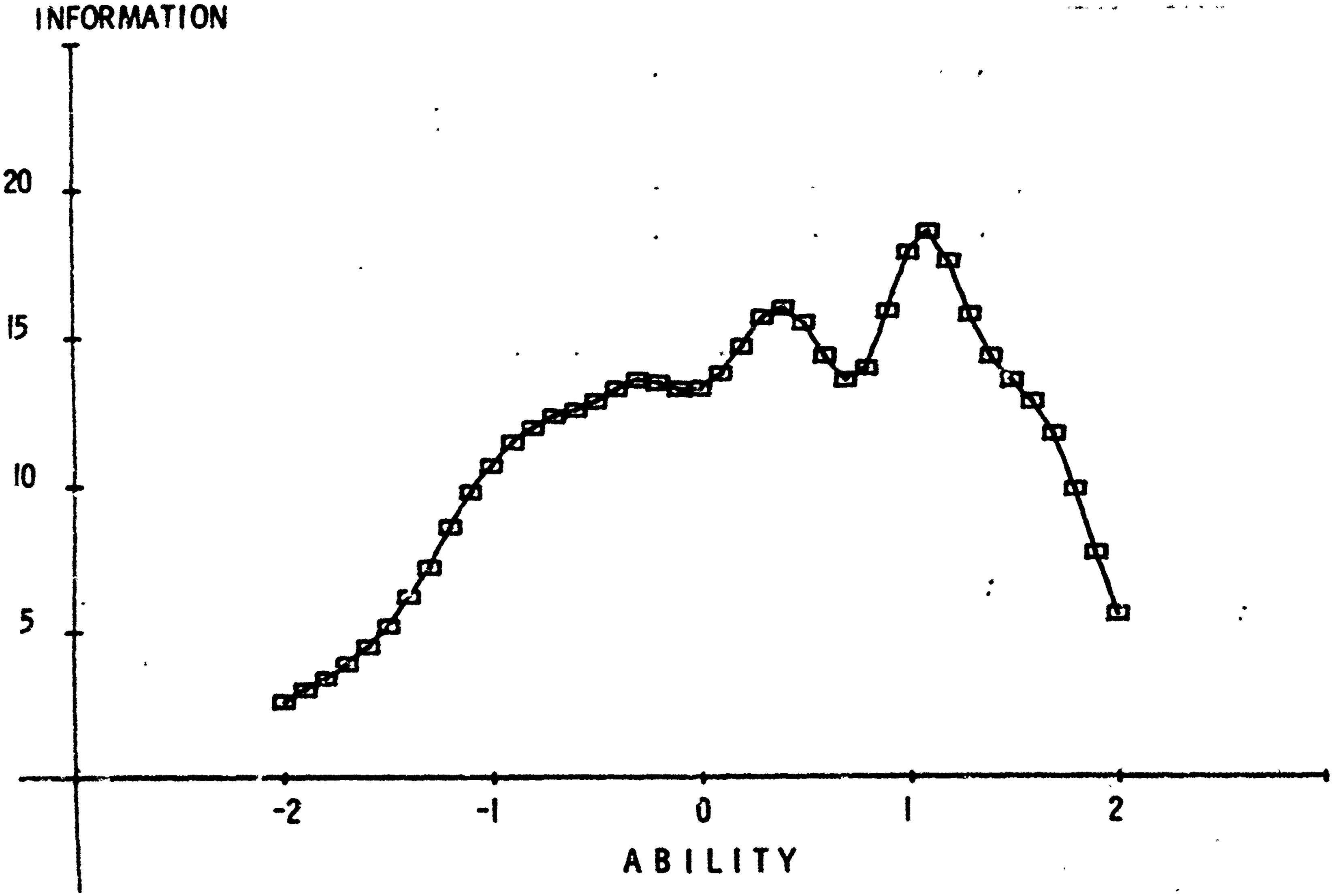


Figure 4. Clusters of test items constituting the six levels of the second-stage multi-level test. Eight arrows indicate the ability estimates resulting from the routing test. Dotted lines indicate the item cluster (level) assigned to the respective routing test ability estimate.

Figure 5. The test information function estimated for the adaptive test of AR.



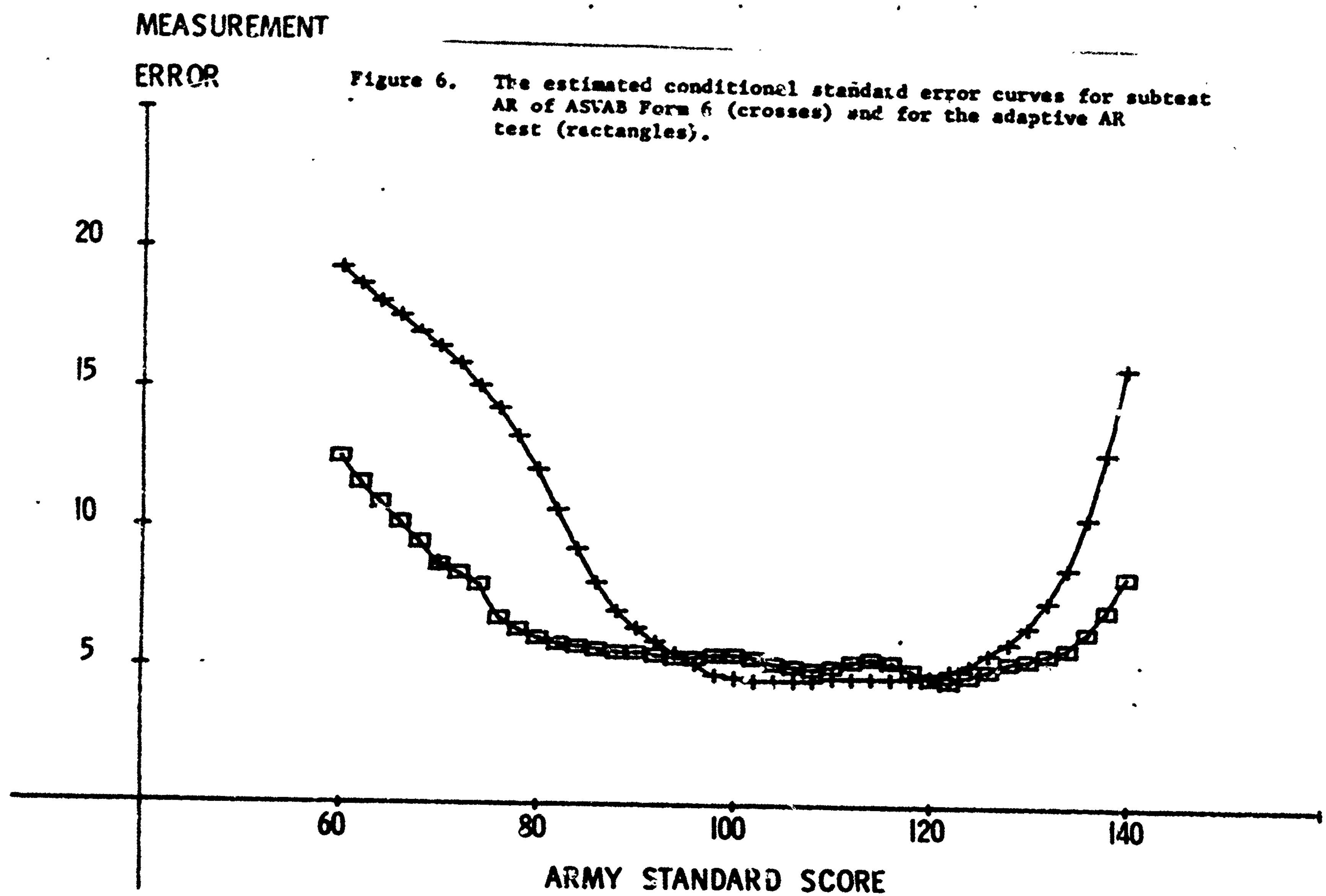


Figure 6. The estimated conditional standard error curves for subtest AR of ASVAB Form 6 (crosses) and for the adaptive AR test (rectangles).

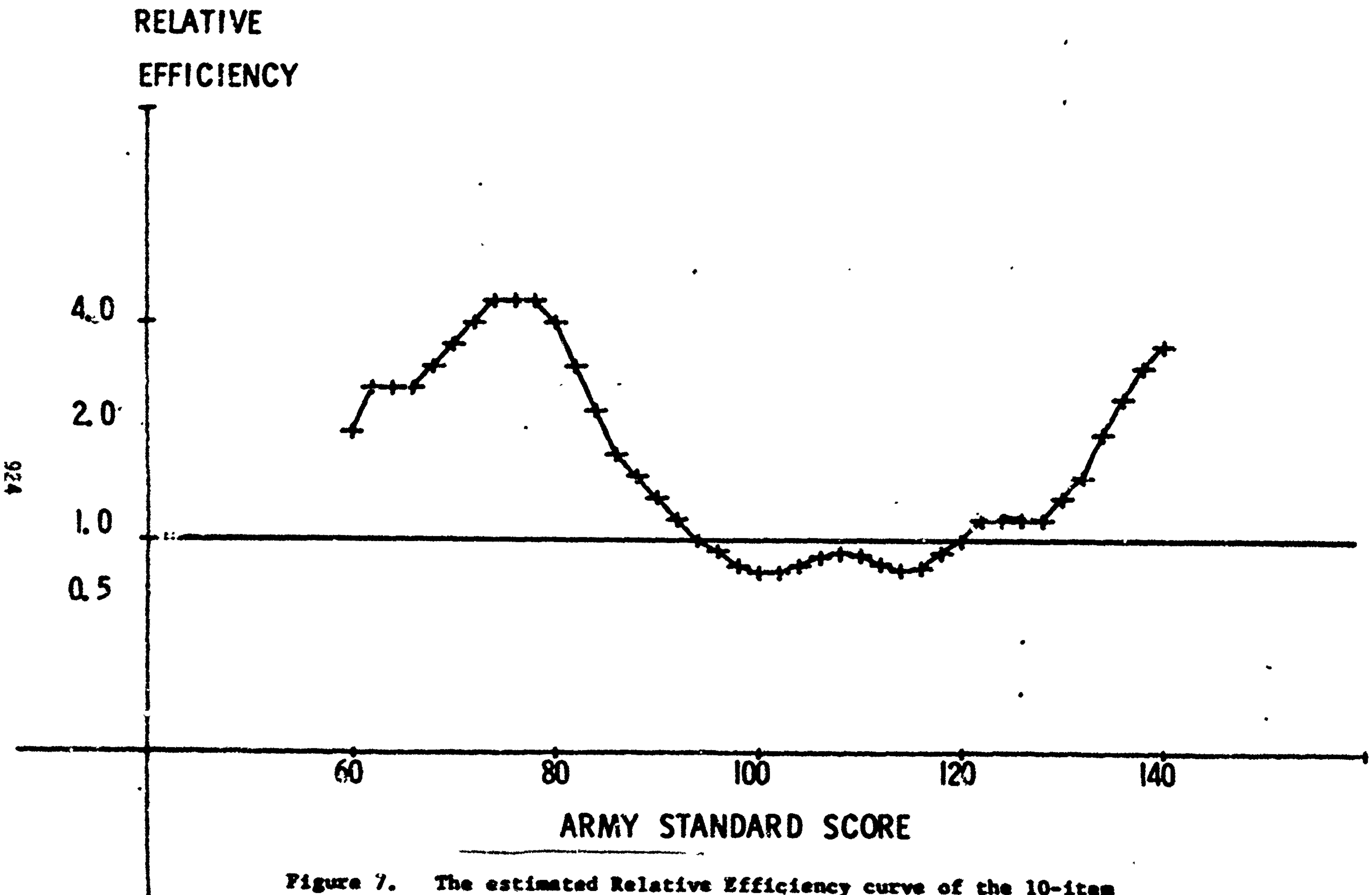


Figure 7. The estimated Relative Efficiency curve of the 10-item adaptive AR test, compared to the conventional AR subtest.

ANALYZING THE TRAINING IMPACT
OF NEW WEAPON SYSTEMS

H. Anthony Baran
Andrew J. Czuchry
Duncan L. Dieterly

ABSTRACT

The objective of the research effort described in this paper is the development of a model that will project a training program for new weapon systems early in their development. It allows for the introduction of time and dollar constraints and uses a basic unit of input defined as a task. The target application is a new concept of avionics integration. However, the computerized analytical model and its associated data bank form the basis of a methodology for allowing the potential training consequences of selected design options for any new weapon system to be more fully considered in the design process.

Quantitative analysis of the potential human resources requirements of weapon systems, still within the design process, has been recently improved by technological advances in the application of simulation modeling techniques. These advances have not, however, been extended to the qualitative aspect of system personnel requirements to enable the simultaneous conduct of a similarly detailed training analysis. This is unfortunate because the analysis of training impacts is essential during the design process to produce designs which maximize operational effectiveness at minimum cost.

The methodology reported addresses the qualitative aspects of human resources requirements through the choice of training options. Using a technique for classifying learning requirements and requisite training options as a function of performance tasks, it bridges design and training in a way which allows cost and operational constraints to be, not only considered but also, traded-off with each other. It provides a capability to rapidly assess training requirements and select a training approach and program most appropriate within the limits established by a set of user specifiable constraining conditions. Examples of these are cost, training time, student flow, maintenance policy, and planned use of job performance aids.

Application of this methodology during the early stages of design can provide a means to relate operational training and personnel constraints and design, early enough to allow designers to incorporate results in the key decisions of design finalization within the systems acquisition process.

INTRODUCTION

The training model described in this presentation is being developed to meet an immediate need for means to assess the impact of a new concept of avionics integration on training. It is also designed to provide a basic tool for examining the consequences of almost any set of circumstances which bears upon training needs and how they are to be fulfilled. Although capable of independent operation, it is a part of a life cycle cost (LCC) modeling system being constructed within a LCC study in the Digital Avionics Information System (DAIS) advanced development program. The overall objective of that study is to assess the LCC impact of the DAIS and also to provide more adequate means for incorporating LCC considerations into design, operations, and maintenance decisions throughout the systems acquisition process, particularly in its early stages.

Although the training model data bank development is currently specific to avionics, the model represents an extremely broad approach to training analysis. Its primary contribution to training technology is its generalizability and the fact that it establishes an increased degree of logic and mechanization in a task which is often thought to be more of an art than a science. Rather than a completed structure, it is the framework of a training evaluation process to be built upon and expanded to more adequately address specific needs.

The model allows a training analyst to assign values to variables describing systems, policy, training operations, resources, and cost. Within the bounds of the user established set of constraints, it produces an estimate of the training program requirement which their interactions generate. Results may be refined by iteratively exercising the model using different values for constraint parameters and/or input data. The means to relate system/policy/resources/cost input data to resultant training impacts are contained in the model.

The rapidity and ease of their exercise, and the way in which the model facilitates their iterative implementation, solves a number of the problems in training impact analysis. Among these are the early identification of excessive requirements, timely investigation of alternatives, and training cost estimation.

The modeling approach to training impact analysis affords a capability which the training community has been seeking for quite some time. It introduces the methodology to play a more active role in controlling the onset of training programs. By allowing attention to be focused on the relative effects of input data changes rather than on the calculations involved in quantifying their interactions, the modeling concept also allows the training analyst to pinpoint causational antecedents in design, policy, etc., which could give rise to problems in the planning of training. This information can be transmitted to the designers of weapon systems and policy planners for their consideration along with other requirements. Providing the capability to impart increased foresight to designers and planners is, in fact, one of the primary functions of the DAIS LCC study under which the model is being developed.

The DAIS advanced development program is an Air Force Avionics Laboratory program seeking to demonstrate a solution to the problems of proliferation and non-standardization of aircraft avionics. It is developing and testing a concept of integrated avionics as an information management system. This concept proposes that the processing, multiplex transfer, and display functions of avionics subsystems be common and serve all the other avionics functions on an integrated basis.

Historically, mission information requirements have been established along essentially autonomous subsystem areas such as navigation, weapon delivery, stores management, and flight control. The resulting complexity of new system configurations, designed to meet these requirements in the fashion in which they were established, has led to ever increasing system support requirements. These translate to increases in system LCC. This, and the fact that nearly all avionics subsystems have trended toward digital methods which can facilitate functional integration across subsystems, were major catalysts of the DAIS program.

Clearly, the DAIS concept possesses a potential to effect significant changes in the design, procurement, operation, and support of weapon systems. Also apparent is the fact that these impacts can be as variable as they are numerous, depending upon such things as degree or manner of implementation. The DAIS LCC study is addressing these issues by expanding upon available technology for identifying and quantifying the consequences of design on system ownership. The DAIS training model and data bank is but one result of a search for means to pinpoint specific impacts of design on individual components of LCC in rapid fashion, using conceptual level design data.

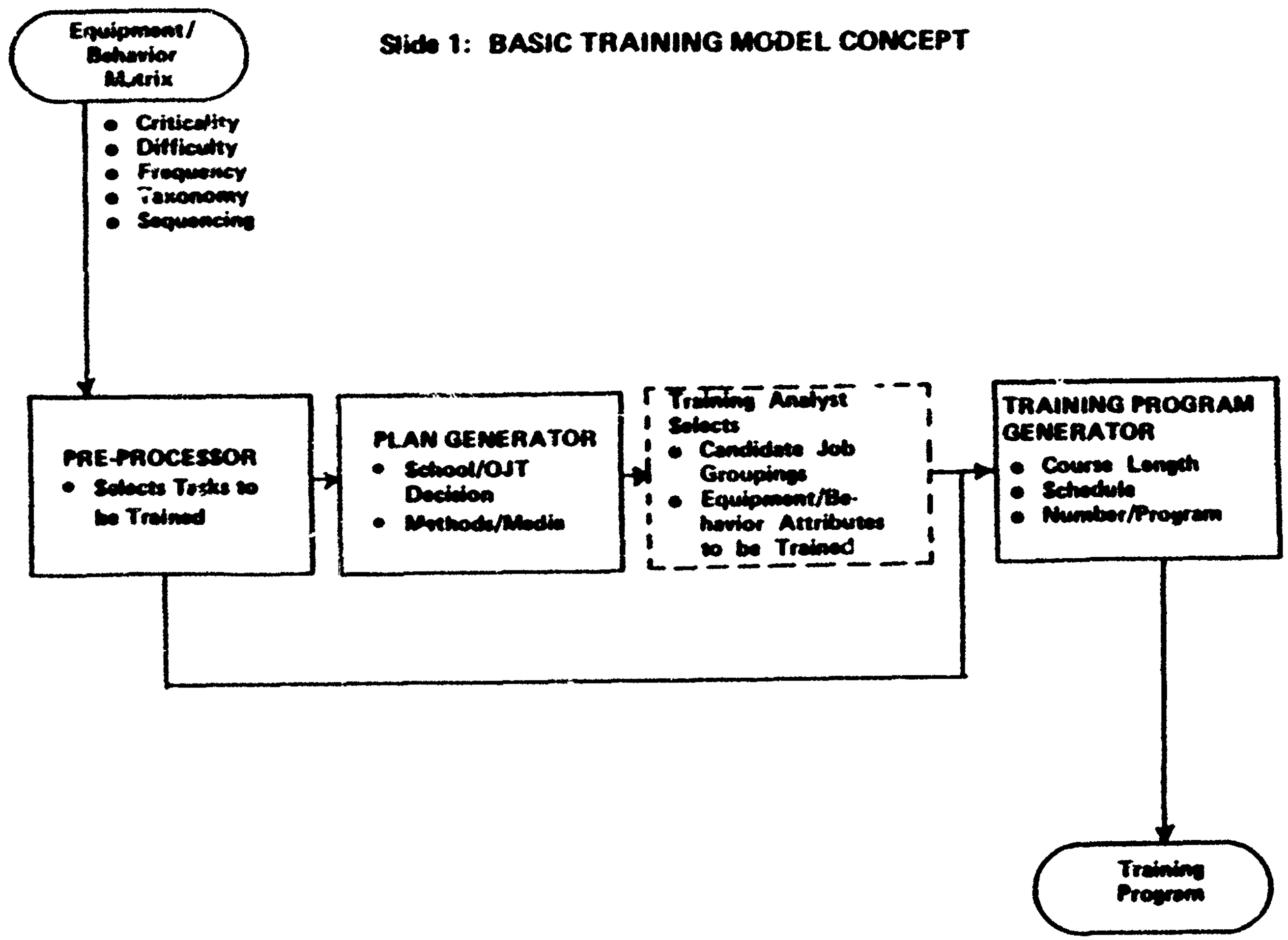
Together, the Air Force Human Resources Laboratory and Dynamics Research Corporation of Wilmington, Massachusetts, are engaged in an effort to construct a LCC modeling system capable of assessing the various impacts of new weapon systems either singly or in concert. One of the components of this system is the DAIS training model. It is a computerized analytical model which, in the context of the overall system, provides requisite information so that DAIS training costs can be computed. This presentation will indicate how it can be used: (1) to assist the training analyst in conducting trade-off studies to define the most cost effective training program; and (2) to suggest a procedure for influencing design and support system concepts based upon associated training requirements and their cost.

As you all probably know, the Instructional System Development process, defined in Air Force Manual 50-2, is the foundation of the Air Force approach to insuring cost effective instruction. The research tool described here parallels that process very closely. It might even be said that it represents a meaningful step toward its mechanization in developing and refining education and training programs. The following is an overview of its operation.

OPERATION OF THE MODEL

The training model (Slide 1) consists of three modules: a pre-processor, and two analytical modules for training plan and training program generation.

Slide 1: BASIC TRAINING MODEL CONCEPT



Operation of the model is predicated upon the establishment of a data bank containing the set of tasks to be learned. Their level or specificity is a user defined variable, allowing for flexibility of task definition. Each task, however, is assigned five descriptor values denoting: frequency, criticality, learning difficulty, taxonomy, and sequencing.

The data bank is inputted to the pre-processor module which screens the total set of tasks, in a series of go no-go decisions, to select those which require training. The selected tasks then become the subset of tasks that are the training requirement. The selection is based upon pre-established descriptor value levels determined by the user. For example, a criteria of tasks of a difficulty level above .60 may be used to discriminate between tasks on the basis of that parameter. Thus, the user maintains control of the decision process by his selection of decision criteria, i.e., parameter combinations and parameter value cut-off points. The list of tasks which the pre-processor determines to be requirements for training retains its associated set of descriptor values and becomes the input data set for the first analytical module which is the training plan generator.

At this point, it is assumed that all of the outputted tasks are to be trained. The user now has the option of designating a value for any one of three constraining conditions: personnel required (number); maximum training cost (dollars); or maximum training time (months). He need, however, only specify the trained personnel requirement to operate the module using internalized data and relationships. The training plan generator then produces an initial training plan. This is a two step process in which a minimum cost school/on-the-job training (OJT) mix is determined, followed by recommendations concerning appropriate methods and media, e.g., lecture, simulation, mockups, actual equipment, etc.

After reviewing the initial training plan, the user may either select a different set of decision criteria and exercise the training plan generator module to obtain another training plan, or continue on to the second analytical module to generate a training program. Generally, the training plan generator will be iterated several times by the user as an investigative/optimization procedure prior to the selection of a training plan to be examined in more detail.

The training program generator uses the outputs of the training plan generator, along with either user specified or standard model values for decision criteria, to produce a representative training program. This consists of a schedule; number of classes per program; number of instructors, simulators, etc., per program; course length; estimated cost; etc. As in the exercising of the training plan generator, it is expected that the user will also use the initial training program as the basis for iteration to examine the effects of changing the values of the input parameters under his control. We will now go into further detail concerning data bank construction and the functions and capabilities of each portion of the model.

TRAINING MODEL DATA BANK

The first step in analyzing the training impact of a new system is the establishment of a data bank containing information for use in translating the equipment and/or maintenance characteristics of that system into the basic elements which govern the establishment of training plans and programs. Basically, this consists of a systems maintenance/operations requirements analysis in terms of tasks and their descriptors, and a subsequent analysis of the identified tasks in terms of the behaviors they subsume. The latter analysis is, in many ways, analogous to the former. It is conducted to achieve a more refined description of tasks in terms of parameters which can later be used to classify and grade them. This classification and grading is the basis used in the pre-processor for decision making concerning which tasks are to be trained and later, in the analytical modeling components, for decisions concerning training plan and program definition. The tasks are then grouped by career field designation. However, this last step is solely for the purpose of data bank organization. It is assumed that each exercise of the model is to be accomplished using tasks within a single personnel category.

Data for the DAIS application was gathered solely in the maintenance area, on the basis of equipment comparability analyses and historical records. It could, however, have been gathered on operator tasks or have been based on survey, interview, or time and motion study techniques. Perhaps one of the most critical aspects of data bank development is the selection of the task descriptors which it will contain. They must achieve a balance between the degree of specificity required to perform meaningful translations and the latitude of applicability required to maintain propriety across a wide variety of tasks. The descriptors we have selected are representative but, not exhaustive.

Each task description consists of graded assessments in terms of the task descriptors used to translate equipment parameters to training parameters. The DAIS training data bank is formatted in the manner of a matrix, as shown in (Slide 2). Tasks are related to equipment and, when appropriate, the maintenance event or operational activity required to restore the equipment to operational readiness. Each task is evaluated in terms of five descriptor parameters: (1) frequency; (2) criticality; (3) learning difficulty; (4) taxonomy groupings; and (5) a parameter describing sequenced tasks to be trained as a group.

Frequency is a relative measure of how often the task occurs or must be performed. Task criticality and learning difficulty are assessed according to Instructional System Development guidelines. Criticality is determined on the basis of two factors: whether the task is required under emergency conditions; and the consequences of inadequate performance. In the training model this is a simple dichotomous choice category. Learning difficulty is assessed on the basis of task complexity and the knowledge and performance requirements associated with it. We are using a five step range to grade the tasks. In the case where tasks can be broken into behaviors, these are graded individually and their scores aggregated to obtain a composite task score.

Slide 2: TRAINING MODEL DATA BANK

Input Data

Task Descriptors Task Identifier	Criticality	Difficulty	Frequency	Psychomotor Level	Cognitive Level	Nesting Para.
Equipment or Event Identifier						

The taxonomy descriptor parameters and the scales which define their levels were adapted from Bloom's representation of human behavior. Tasks are described in terms of two classifications: cognitive and psycho-motor activity. The tasks are then judged within each category according to the scalar level designates shown in (Slide 3). For example, the behavior of coding, within the task of computer programming, might be a cognitive level four behavior and a psycho-motor one behavior; whereas the behavior of soldering, within the task of equipment reassembly, might be a cognitive level two and a psycho-motor level three.

It might be well to note at this point that we recognize that this particular method of task classification and grading is, by no means, comprehensive and that the scaling is subjective. However, they are practical starting points for providing a common denominator for a wide variety of tasks. This is an essential ingredient in translating equipment and/or maintenance characteristics of a system into the task performance criteria which determine a need for training programs.

The fifth descriptor parameter, the task sequencing or "nesting" factor, is relatively self-explanatory. It serves to designate tasks which logically fall together, either on the basis of their performance interaction or requirements generated by the actual provision of training.

TRAINING MODEL PRE-PROCESSOR

The pre-processor contains a set of selection algorithms which determine the subset of tasks that become the training requirements. The five descriptor parameters are the criteria for a series of decision filters, whose cut-off levels are set by the user. Their function varies from simple sorting, as shown in (Slide 4), to weighted averaging for the calculation of a task parameter which we call task intensity. This combinatorial parameter can be used in conjunction with an algorithm within the module to more closely examine the potential effects of either operational or policy alternatives of the system which will provide the training. This mode of operation allows the user great flexibility in emphasizing or de-emphasizing any or all of the five descriptor parameters.

Slide 3: BASIC TAXONOMY

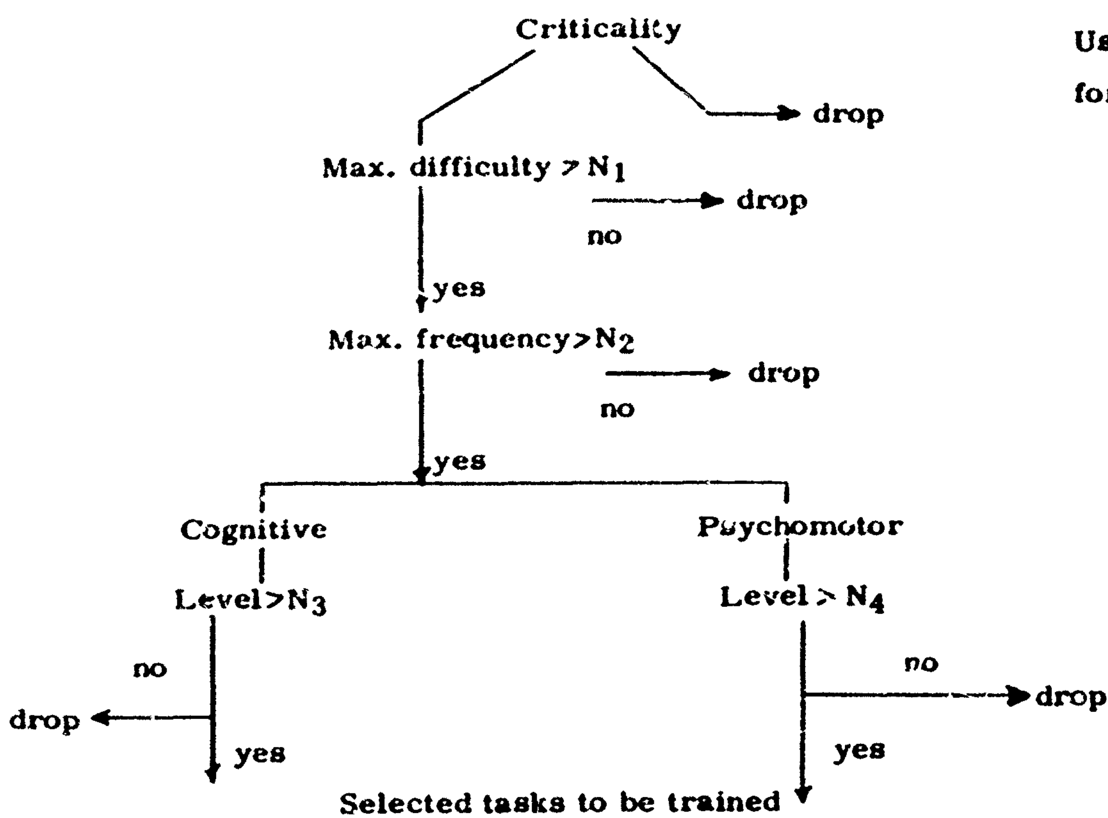
Cognitive		Psychomotor	
5	Evaluate	5	Naturalization
4	Synthesis	4	Articulation
3	Analysis	3	Precision
2	Application	2	Manipulation
1	Comprehension	1	Initiative

EXAMPLE

Computer Programming	*4*	*Cognitive*
Coding	*1*	*Psycho-motor*

Slide 4

EXAMPLE DECISION ALGORITHM FOR TRAINING MODEL PRE-PROCESSOR



User specifies values for N_1, N_2, N_3, & N_4.

In order to assess the impact of various criteria on the establishment of training requirements, the user inputs his selection of descriptor parameter cut-off levels (decision criteria), and operates the pre-processor to yield a set of training requirements and their characterizations in terms of the descriptor parameters. Each time the training analyst exercises the pre-processor with a different set of decision criteria, the result is a different set of training requirements. Thus he can analyze the direct impacts of design and/or policy on training requirements.

TRAINING PLAN GENERATOR

The training plan generator allows the user to analyze the direct impacts of varying either a personnel quantity, training time, or cost constraint on the establishment of a training plan. Its required inputs are the outputs of the pre-processor and specification by the user of the number of trained personnel required. The opportunity to include cost and time constraints and parameters which reflect his knowledge of such things as training facilities, costing procedures, personnel categorization within the organization, etc., allow him to tailor the input to reflect known constraints governing the capabilities of the training providers or conduct sensitivity analyses.

The output of the training plan generator module is a minimum cost initial training plan that meets the inputted constraints. It consists of task groups identified by a school/OJT mix and appropriate training methods and media, along with cost and time estimates. This is accomplished in a two step process which first determines an optimum school/OJT mix and then procedes to select appropriate training methods and media. Both steps are accomplished on the basis of the descriptor parameter values assigned to each task. To determine the school/OJT mix, the module takes the training requirement tasks, along with the user input of number of personnel required, and examines their descriptor values in terms of a number of equations relating cost, time, and type of training.

Standard values within the module for the descriptor parameter decision criteria, as well as for factors within the equations, are addressable by the user for revision. It should also be noted that an assumption is made that school and OJT are equally effective, at least in rendering satisfactory results in terms of trainee proficiency. It is also assumed that all training is to be task oriented. Recognizing that these assumptions greatly simplify the job of the training plan generator, we have allowed in the module for their revision.

Also, while the cost relationship equations within the module appear sufficient for the school/OJT decision, they are not used in the estimation of the cost of the training program which the model will generate as a final product. Detailed costing of the training program is a separate effort which uses the cost figures generated by the training model as a starting point.

At this point, the module has grouped the tasks to be trained on the basis of the sequencing descriptor parameter, and designated them by training type. In the second step of the training plan process, they are again sorted by means of an algorithm which is essentially a task/training objective comparator. It first maps a training objective profile for each task group, on the basis of the taxonomy descriptor parameter values for the tasks within each group, (Slide 5), and compares the results with similar profiles established for various training methods and media. The profile of training objectives is the common denominator for tasks and training methods and media. Those established for the training methods and media are based on criteria established by Parker and Down in 1961 (Slide 6). Once the task groups have been associated with specific training methods and media, an output is generated which provides a training plan broken out by task group, by school/OJT designation, and by methods/media recommendation. Cost estimates and time requirements are also included (Slide 7).

The structure of the training plan generator allows the user considerable discretion in its control. In similar fashion to his control over the training type decision process, he can alter the rules for the process which associates tasks with training objectives or simply specify a particular mapping. He can also alter the rules for the process which associates training objectives with training methods and

Slide 5

TASK/TRAINING OBJECTIVE MAPPING

Taxonomic Description	Training Objective
Psychomotor, Imitation Cognitive, Comprehension	Learning Identifications
Psychomotor, Manipulation Psychomotor, Precision	Learning Perceptual Discrimination
Psychomotor, Articulation Cognitive, Application	Understanding Principles and Relationships
Cognitive, Analysis Cognitive, Synthesis	Learning Procedural Sequence
Cognitive, Evaluation	Making Decisions
Psychomotor, Naturalization	Performing Skilled Perceptual Motor Acts

Slide 6

ALGORITHMS FOR SELECTING TRAINING METHODS AND MEDIA

Training Objectives	Training Type	Method/Media
Learning Identifications	School	Discussion/Transparencies
	OJT	Informal Lecture/ Transparencies
Learning Perceptual Discrimination	School	Simulation/Training Film
	OJT	Demonstration/ Training Film
Understanding Principles and Relationships	School	Simulation/Simulator
	OJT	Performance/Mock-ups
Learning Procedural Sequence	School	Performance/Simulator
	OJT	Performance/Training
Making Decisions	School	Simulation/Simulator
	OJT	Performance/Training Films
Performing Skilled Perceptual Motor Acts	School	Performance/Simulator
	OJT	Performance/Actual Equipment

Slide 7: TRAINING PLAN GENERATOR OUTPUT

Group j - Sys. i	Mode	Time	Method	Medium
AGE F/L Subtask - 121 Subtask - 122 Subtask - 123	School	4 Weeks	Discussion	Transparencies
Subtask - 151	OJT	8 Weeks	Performance	Mock-ups
TS F/L Subtask - 231 Subtask - 232 Subtask - 233 Subtask - 234	School	6 Weeks	Simulation	Training Film
R&R Subtask - 321	OJT	4 Weeks	Demonstration	Training Film
Subtask - 361 Subtask - 362 Subtask - 363	OJT	10 Weeks	Performance	Actual Equipment
. . . SHOP–N Subtask - 031 Subtask - 032	School	8 Weeks	Performance	Simulator

Group j - Sys. i+1	Mode	Time	Method	Medium
AGE F/L Subtask - 151				

media. These capabilities are available as input selections and do not require any re-programming. It should be noted that, while the training methods/media selection mechanism does not break down recommendations to specific implements nor deal with specific numbers of instructors, simulator types, etc., it is capable of doing so. The obstacle to this finer grained analysis is a lack of data at the present time, not a programming limitation.

It should be clear at this time that each successive phase of the model's exercise allows the user to investigate the impacts of results that have preceded it. By inputting precise data, as it becomes known, the user can generate very precise impact estimates. By using the standard relationships within the model itself, he can also obtain relative impact estimates of great value early in the system design process. Training program generation is the last phase of modeling activity. It provides information needed to calculate a cost estimate for the training plan selected and/or optimize resource consumption.

TRAINING PROGRAM GENERATOR

The training program generator module takes the outputs of the training plan generator and produces a training program on the basis of internalized rules of resource management. The training program consists of a schedule, number of students per program, number of instructors, course length, etc. The user specifies the required number of trained personnel needed per year, minimum/maximum class size, resources available and cost, and a ranking of both resources and training program objectives according to their relative importance. For resources, this ranking is usually based on availability and/or cost. Training program objectives may be ranked on almost any basis such as safety, mission, or performance requirements. In the event that the user chooses not to specify a ranking, the algorithm within the training program generator assumes that simulators are the high cost drivers and that instructors are the scarcest resource. Given the training plan, the ranking of resources and training objectives, and the personnel requirements, etc., the algorithm undertakes an optimization routine to yield a reasonable first cut at a training program of sufficient detail to be costed (Slide 8).

Slide 8: TRAINING PROGRAM GENERATOR OUTPUT

COURSE SCHEDULE – WEEKS: 2, 10, 20, 30, 40, 50, 60

Group j Subsystem i							
Class 1 (18)	Task 12 TTS TRAN	Task 15 OJT MOCK	Task 23 TTS FILM	Task 32 OJT FILM	Task 36 OJT EQIP	. . . .	Task 03 TTS "SIML"
Class 2 (18)	Task 03 TTS "SIML"	Task 12 TTS TRAN	Task 15 OJT MOCK	Task 23 TTS FILM	Task 32 OJT FILM	Task 36 OJT EQIP	. . . . . .
Class 3 (18)	Task 12 TTS TRAN	Task 15 OJT MOCK	Task 03 TTS "SIML"	Task 23 TTS FILM	Task 32 OJT FILM	Task 36 OJT EQIP	. . . . . .
Class 4 (17)	Task 12 TTS TRAN	Task 15 OJT MOCK	Task 23 TTS FILM	Task 32 OJT FILM	Task 03 TTS "SIML"	Task 36 OJT EQIP	. . . . . .
Class 5 (17)	Task 12 TTS TRAN	Task 15 OJT MOCK	Task 23 TTS FILM	Task 32 OJT FILM	Task 36 OJT EQIP	Task 03 TTS "SIML"	. . . . . .

Group j Subsystem i+1	

Results may be iterated to determine various sensitivities. Doing so may reveal inordinacies in resource consumption, cost, etc., which might be avoided by changes upstream closer to the equipment design end of the training analysis procedure. The capability for iteration, within and across the components of the training model, using different sets of criteria is, in fact, its strongest feature.

SUMMARY

In summation (Slide 9), I would like to re-emphasize that the training model presented today should be thought of as the first of a series to be modified and refined in the future. Perhaps its most significant aspect is its potential to be useful in an extremely diversified array of applications. This potential is based upon its generalizable structure. This is best summed up by saying that it provides a means to standardize training impact analysis.

An extensive repertoire of training technology exists which supports the design of training systems. The training model provides a means to facilitate its concerted and timely application. Decisions concerning the establishment of training plans and programs are becoming more and more difficult due to the ever increasing number of variables which training analysts must consider. Many of these are incidental to training objectives. I refer to such things as cost, lead time, and other variables which are primarily external to the analysis of training.

Despite a good technology base, the increasing number of variables and the tracking of their interactions tend to preclude comprehensive analysis by the sheer complexity of their calculation. This situation is further exacerbated by the narrowness of the time-frame in which the results of training analyses may provide useful feedback to designers and planners. It is also assuming increased importance as planners become more attentive to the life cycle cost aspect of systems acquisition.

Slide 9

TRAINING MODEL SUMMARY

- PROVIDES A MEANS TO STANDARDIZE TRAINING IMPACT ANALYSIS
- FACILITATES APPLICATION OF TRAINING TECHNOLOGY
- ENCOURAGES PARTICIPATION IN WEAPON SYSTEM DESIGN PROCESS
- OFFERS POTENTIAL COST AVOIDANCE THROUGH EARLY ANALYSIS OF TRAINING IMPACTS

Training is expensive. Its expense reaches far beyond the cost of producing trained personnel. The real cost of training includes penalties paid in terms of lost opportunities. These are the costs associated with failures to capitalize on numerous potentials for cost avoidance, due to an inability to extend the analysis of training requirements beyond its present role of reacting to given sets of conditions. Clearly, it would be advantageous for training analysis to change from a post hoc activity to become an integral part of the weapon system design process. This requires a capability to take part in decisions concerning aspects of design and policy which create training requirements. The fulfillment of training requirements is only half the battle.

The modeling approach to training impact analysis can make this change possible. It can increase the speed and systematization of the procedures entailed in training planning and resource management. Quite apart from its potential to aid designers in developing more maintainable and cost effective systems, its versatility makes it ideal for even the most mundane problems concerning the provision of training and resource management.

The training model described is a first step in defining a methodology for the practical application of the modeling approach. What remains is for the training community to continue its development in terms of data and criteria. The model itself stands alone as a mechanism capable of performing many of the required data manipulations entailed in training impact analysis, allowing the user to quickly tradeoff alternatives.

REFERENCES

1. Czuchry, Andrew J.; Engel, Herbert E.; Dowd, Richard A.; Baran, H. Anthony; Dieterly, Major Duncan L.; and Greene, Ron. Mid-1980s Digital Avionics Information System Conceptual Design Configuration, AFHRL-TR-76-59, AD-A032137. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resource Laboratory, May 1976.

2. Engel, Herbert E.; Glasier, John M.; Dowd, Richard A.; Bristol, Marjorie A.; Baran, H. Anthony; and Dieterly, Major Duncan L. Digital Avionics Information System (DAIS): Current Maintenance Task Analysis. AFHRL-TR-76-71, AD-A035683. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, October 1976.

3. Czuchry, Andrew J.; Engel, Herbert E.; Bristol, Marjorie A.; Glasier, John M.; Baran, H. Anthony; Dieterly, Major Duncan L. Digital Avionics Information System (DAIS): Mid-1980s Maintenance Task Analysis. AFHRL-TR-77-45. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, July 1977.

4. Parker, James F.; and Downs, J. E. Selection of Training Media. ASD-TR-61-473. Wright-Patterson AFB, OH: Aeronautical Systems Division, Aerospace Medical Laboratory, Behavioral Sciences Laboratory, September 1961.

5. Bloom, B.S., Taxonomy of Educational Objectives: Classification of Education Goals, Cognitive and Affective Domains. McKay Publishers, 3rd Ave., N.Y., N.Y., 1956.

6. Department of the Air Force. Instructional System Development. Air Force Manual, AF Manual 50-2, July 1975.

CONCERNING THE AUTHORS

Mr. Baran is the DAIS Life Cycle Costing Study Manager. He is a civilian employee of the Advanced Systems Division of the Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, Ohio. Major Dieterly, of the same organization, is a Deputy Program Manager (Human Resources) of the DAIS advanced development program. Dr. Czuchry is the director of the Advanced Systems Department of Dynamics Research Corporation, Wilmington, Mass., the contractor for the DAIS Life Cycle Costing Study.

Analysis of Flight Clothing Effects on Aircrew Station Geometry

Lt Cmder Harvey G. Gregoire

INTRODUCTION

BACKGROUND

1. Typically aircrew station geometry requirements have been based on nude male anthropometric data taken from measurements on a standard anthropometric chair, a flat seat with a 90 deg perpendicular back surface. Since aircrew persons do not fly nude, nor do they sit on a flat surface with a 90 deg perpendicular back, nor are they all male anymore, it is necessary to quantify the effect of those items worn in the aircrew station environment. The necessity to quantify the effects of personal flight clothing and equipment is particularly important in presently developing tactical aircraft since the anticipated higher g operational environments are more restrictive to anthropometric mobility than earlier models of tactical aircraft. Additionally, the primary flight instrument status of Heads-Up Displays and similar electro-optical devices may limit the design eye reference of the pilot's eye position to a greater degree than other similar aircraft models.

2. Many of the prior research efforts in the area of quantifyng the effects of flight clothing relative to anthropometric accommodation have generally been item specific; i.e., the effects of wearing a pressure suit or a helmet, etc. There has been little research, if any, on the anthropometric effects of an entire complement of flight clothing and equipment.

3. Military Standard 1472B, the Human Engineering Design Criteria for Military Systems, Equipment, and Facilities, specifies that suitable allowances must be made for the design-critical dimensions imposed by protective clothing or equipment. Providing "suitable allowances" for an unknown quantity can be difficult at best, if not impossible. The failure to use data concerning the effect of flight clothing and equipment on anthropometry in the design of aircrew stations has historically been costly in terms of aircrew safety, efficiency, mobility, and comfort.

4. The specific goal of this analysis was to provide data to quantify and describe the effect of increased bulk and decreased mobility resulting from the wearing of summer and winter flight clothing and equipment in a typical ejection seat environment.

5. The data derived from this evaluation can be used in the following applications: (1) as correction constants to be applied to current computer based simulation models which have as their goal the early (blueprint) detection of inconsistencies between planned cockpit geometry and anthropometric characteristics of the intended user population, (2) as a design aid to engineers tasked with providing the anthropometric accommodation in aircrew stations specified by military standards, and (3) as a reference aid to those organizations tasked with developing aircrew clothing and equipment possessing the minimum bulk, weight, and mobility restriction commensurate with the necessary protective characteristics.

DESCRIPTION OF TEST FACILITY

6. Comparative anthropometric measurements of subjects in unclad, summer flight gear and winter flight gear configurations were made using a Navy 64A105H1-1 Integrated Measuring Anthropometric Device and a standard medical weight scale.

7. The cockpit specific anthropometric range of motion measurements was made in a Douglas ESCAPAC IF-3 ejection seat and restraint system modified with adjustable point-of-reference protractors positioned at range-of-motion joints (i.e., neck, clavicle, elbow, wrist, lumbar, hip, and ankle areas). The ejection seat selected was typical of lap belt and inertia-reel torso restraint systems found in ejection-seat equipped tactical aircraft.

8. This evaluation investigated the flight clothing and equipment effects on volume and mobility for a sample of aircrewman representative of the entire spectrum of Naval aviator body sizes. The 1964 Anthropometry of Navy Aviators Survey, which listed body size data for 96 measurements of 1,549 aviators, was used for anthropometric percentile-rank criterion of the measurements evaluated except for buttock-leg dimensions. A 1976 data sample compiled on anthropometric variables for 969 aviators was used to define the buttock-leg percentile-rank criterion for this evaluation.

9. The anthropometric dimensions, joints, and respective range-of-motion measurements included:

a. Dimensions.

(1) Weight.

(2) Stature.

(3) Standing waist height.

(4) Functional arm reach.

(5) Shoulder-elbow length.

(6) Forearm-hand length.

(7) Hand length.

(8) Standing hip breadth.

(9) Sitting height.

(10) Bideltoid diameter.

(11) Buttock-knee length.

(12) Sitting hip breadth.

(13) Popliteal height.

(14) Buttock-leg length.

(15) Foot length.

b. Joints and respective ranges of motion.

(1) Neck - head/look angle.

(a) Elevation.

(b) Declination.

(c) Azimuth right.

(d) Azimuth left.

(2) Clavicle/humeral - extended arm movement.

(a) Elevation.

(b) Declination.

(c) Azimuth right.

(d) Azimuth left.

(3) Elbow - lower arm movement (measured with upper arm extended horizontally and vertically from clavicle joint).

(a) Elevation.

(b) Declination.

(c) Azimuth.

(4) Wrist - extended hand movement.

(a) Elevation.

(b) Declination.

(c) Azimuth right.

(d) Azimuth left.

(5) Lumbar - torso movement, sitting.

(a) Declination.

(b) Torsion right.

(c) Torsion left.

(6) Hip - upper leg movement, sitting.

(a) Elevation.

(b) Azimuth right.

(c) Azimuth left.

(7) Knee - tibial movement.

(a) Elevation.

(b) Declination.

(8) Ankle - foot movement.

(a) Elevation.

(b) Declination.

(c) Azimuth right.

(d) Azimuth left.

10. The parameters for both series of anthropometric dimensions and angular range-of-motion joints were selected from a crew station assessment of reach computer based simulation model. Over 2,300 measurements were taken for this evaluation.

11. The scope of the flight clothing and equipment evaluated included those current inventory items typically worn by those Navy crewmen who fly tactical and training aircraft equipped with ejection seats.

12. With the exception of those data directly affected by the torso harness and ejection seat restraint systems, other data can be applicable to nonejection seat aircraft.

METHOD OF TESTS

13. The subject crewmen were measured in three separate configurations: (1) unclad, (2) dressed and equipped for summer flight, and (3) dressed and equipped for winter flight. Each dimensional and angular measurement was made four times and averaged to reduce measurement error variability. The quantification procedures are listed below:

a. The subject was weighed.

b. Cockpit specific anthropometric measures were made using the Navy 64A105H1-1 Integrated Anthropometric Measuring Device. Data were recorded on an anthropometric data form.

c. The subject was seated in the ejection seat. Specially mounted transparent protractors were then adjusted horizontally or vertically with the protractor center of radius point aligned with the estimated locus of the joint center of mass. The protractor zero deg reference line was then adjusted vertically and horizontally forward from the subject's respective joint. The subject then moved his joint segment (e.g., arm around clavicle joint) to a point of maximum possible elevation, declination, or azimuth. The experimenter aligned an index marker line which originated in the protractor center of radius with the estimated midline of the respective segment and read the degrees of rotation from zero deg as indicated on the protractor by the index marker line. The maximum angles of motion about joints were recorded on a second anthropometric data form.

d. Additionally, while secured to the ejection seat lap belt and inertia-reel torso restraint system, each subject's reach distance was measured relative to three specified "reach zones." Zone 1 defines the subject relaxed in a locked harness reaching to controls without straining against the harness. In Zone 1, the lumbar, thoracic, interclavicular, and clavicular segments do not move. In Zone 2, the subject strains against the locked harness to obtain maximum reach. The lumbar, thoracic, and interclavicular segments do not move except for the stretch in torso restraint system. The clavicular segment does move since it is not securely held by the torso harness and restraint system. In Zone 3, the shoulder harness is unlocked and the subject is free to lean forward or to the side to obtain maximum reach within the limits of shoulder harness strap length. The lumbar and thoracic segments move within the limits of shoulder harness strap length. The reach distances were measured from the thumb and forefinger grasp to a point at the intersection of the seat back surface and top surface midpoint of the subject's shoulder.

RESULTS AND DISCUSSION

14. The subjects used in this evaluation were seven males, carefully selected to represent the range of anthropometric characteristics found in the Naval aviation population. Subjects representative of 5th, 25th, 50th, 75th, and 99th percentile population members relative to stature and weight were selected. For subjects 1 through 5, each of the 16 anthropometric variables was screened to be within one standard deviation of the population percentile equivalent being represented.

15. The primary purpose of the evaluation was to quantify the added bulk, displacement of posture, and restriction of mobility which results from the average effects of flight clothing and equipment. Therefore, population-wide representative sampling of pertinent anthropometric parameters was employed. The data are, therefore, presented as plus or minus correction factors relative to the dimensional and angle of motion differences quantified between unclad and summer gear and between unclad and winter gear configurations. The average increased bulk anthropometric dimensional correction factor data are presented in appendix A.

16. For angular quantification, a forward-facing seated posture was assumed by the subjects. All joint measurements were made on the right side of the body; left side mirror-image reciprocals were assumed. Vertical measurements were made from a line extending 90 deg to the right of the joint at zero deg elevation. All horizontal measurements were from a line extending forward of the joint at zero deg azimuth. The angular quantifications of average decreased mobility resulting from summer and winter flight gear with locked torso restraint systems are presented in appendix B. Appendix C presents reach data as a function of reach zone and flight gear worn.

RECOMMENDATIONS

17. A maximum effort redesign of the complete flight clothing and equipment system is necessary to reduce the bulk and weight effects of such clothing and equipment on mobility within an aircrew station.

18. When designing crew station geometry and locating controls and displays, designers should incorporate the maximum available data describing reduction in anthropometric mobility and increase in anthropometric volume resulting from flight clothing and equipment worn on the body.

19. The following comments are relative to bulk and mobility restrictions per item or per group of items comprising the flight clothing and equipment.

a. Helmet (APH 6-3)/Oxygen Mask (A13-A) - Five and one-half lb (2.5 kg); weight, bulk, and oxygen hose/regulator "drag" compromise vertical and horizontal head motion and look angle. The anti-exposure suit hampers horizontal mobility less than it does vertical mobility.

b. Flying coveralls (CSFRP-1), gloves (GS1FRP-1), torso harness (MA-2) - Six and six-tenths lb (3.0 kg); weight and bulk not oppressive. When secured to lap belt and shoulder restraint, mobility is naturally restricted. However, redesign of the lap belt to an inertia system such as the shoulder restraints and increasing shoulder inertia-reel strap length would ease mobility in Zone 3 conditions. The flight gloves were the least bulky and least restrictive item of wear.

c. Anti-G coveralls (MK-2A) - Two and two-tenths lb (1.0 kg); slightly restrictive due to necessary tight fit. As a result of interviewing operational pilots, it was determined that this item was generally not accepted to wear in conjunction with CWU-33P anti-exposure suit.

d. Survival vest (SV-2A) - Two and four-tenths lb (1.1 kg); weight and bulk interfere with torso and arm movements.

e. Boots (B 21408) - Four and five-tenths lb (2.0 kg); slight mobility restriction due to weight and length of vertical dimension.

f. Life preserver (LPA-2) - Four and five-tenths lb (2.0 kg); displaces posture slightly due to packaging. Occasional interference with inertia-reel shoulder straps.

g. Anti-exposure suit (CWU-33P) - Six lb (2.7 kg); this was by far the bulkiest, most restrictive item of equipment. The anti-exposure suit significantly reduced angle of motion in the arms, legs, and torso. The bulk was restrictive not only about the shoulders, elbows, and knees, but increased the effective retention of the torso system regardless of harness locked or unlocked condition. Reach to cross-cockpit, vertical, and side-console areas was considerably hampered, if not prevented, by the anti-exposure suit. Some subjects had difficulty reaching the overhead face-curtain ejection handle as a result of the anti-exposure suit bulk and mobility restrictions.

h. The total weight of either summer or winter gear was subjectively identified as one of the more objectional factors of the flight clothing and equipment by each of the subjects as well as numerous aircrewmen interviewed during the project.

i. All aircrewmen involved in the project expressed the need for an all-encompassing integrated redesign of the entire package of personal flight equipment which would reduce weight and increase mobility.

SUMMARY TABLE OF AVERAGE FLIGHT
CLOTHING/EQUIPMENT DIMENSIONAL CORRECTION FACTORS

Anthropometric Measurements	Mean Differences Between Nude Dimensions and Summer Flight Gear	Mean Differences Between Nude Dimensions and Winter Flight Gear
1. Weight	+28.3 lb (+12.8 kg)	+32.0 lb (+14.5 kg)
2. Stature	+3.2 in. (+8.1 cm)	+3.2 in. (+8.1 cm)
3. Waist height	+1.2 in. (+3.1 cm)	+1.2 in. (+3.1 cm)
4. Arm reach (1)	+.3 in. (+.8 cm)	+.5 in. (+1.3 cm)
5. Shoulder-elbow length	+.1 in. (+.3 cm)	+.6 in. (+1.5 cm)
6. Forearm-hand length	+.1 in. (+.3 cm)	+.3 in. (+.8 cm)
7. Hand length	0	0
8. Hip breadth, standing	+1.1 in. (+2.8 cm)	+1.5 in. (+3.8 cm)
9. Sitting Height (2)	+2.2 in. (+5.6 cm)	+2.5 in. (+6.2 cm)
10. Eye height, sitting	+.3 in. (+.8 cm)	+.5 in. (+1.3 cm)
11. Bideltoid diameter	+.2 in. (+.5 cm)	+1.8 in. (+4.6 cm)
12. Buttock-knee length	+.2 in. (+.5 cm)	+.4 in. (+1.0 cm)
13. Hip breadth, sitting	+.9 in. (+2.3 cm)	+1.8 in. (+4.6 cm)
14. Popliteal height, sitting	+.2 in. (+.5 cm)	-.1 in. (-.3 cm)
15. Buttock-leg length (3)	+1.4 in. (+3.6 cm)	+1.7 in. (+4.3 cm)
16. Foot length	+1.4 in. (+3.6 cm)	+1.4 in. (+3.6 cm)

NOTES: (1) Clavicular joint, humeral, radial, hand finger-grip links.
(2) Lumbar, thoracic, vertical neck, lower head, upper head links.
(3) Femoral, Tibial foot links.

SUMMARY TABLE OF AVERAGE FLIGHT CLOTHING/EQUIPMENT CORRECTION FACTORS FOR JOINT-MOTION REDUCTION(1) ANGULAR DIFFERENCE DATA

Joint		No Flight Gear Average	Differences in Summer Flight Gear	Differences in Winter Flight Gear
Neck:	elevation	73	-22	-34
	declination	61	-12	-16
	azimuth-right	85	-8	-18
	azimuth-left	85	-8	-18
Arm:	elevation	105	-22	-47
	declination	152	-13	-26
	azimuth-right	132	-1	-12
	azimuth-left	55	-19	-38
Elbow:	elevation	116	-22	-33
	declination	72	-7	-13
	azimuth-left	63	-8	-14
Wrist:	elevation	61	-1	-3
	declination	75	-1	-11
	azimuth-right	44	0	0
	azimuth-left	26	0	0
Torso:	declination(2)	86	-55	-68
	torsion-right	45	-14	-32
	torsion-left	45	-14	-32
Leg: (femur)	elevation	46	-8	-20
	azimuth-right	7	-2	-3
	azimuth-left	28	-6	-11
Ankle:	elevation	23	-9	-9
	declination	15	-4	-5
	azimuth-right	45	-8	-8
	azimuth-left	40	-9	-10

NOTES: (1) Measured in degrees. Corrections are ± from right arm and leg extremities, left side mirror image is assumed.

(2) Average of lumbar and thoracic link harness unlocked.

SUMMARY TABLE OF AVERAGE FLIGHT CLOTHING/EQUIPMENT CORRECTION FACTORS REACH ZONE DATA FOR JOINT-MOTION REDUCTION

	Zone 1	Zone 2	Zone 3
Summer Gear	32.1 in. (81.5 cm)	36.8 in. (93.5 cm)	43.5 in. (110.5 cm)
Winter Gear	32.2 in. (81.8 cm)	35.3 in. (89.7 cm)	40.3 in. (102.4 cm)

ANALYSIS OF QUESTIONNAIRE DATA TO IDENTIFY "ACE" ATTACK HELICOPTER PILOTS

Robert F. Eastman
Marie Leger
Brian D. Shipley, Jr.

US Army Research Institute Field Unit
Fort Rucker, Alabama 36362

At the request of US Army Training and Doctrine Command, the Army Research Institute Fort Rucker Field Unit is engaged in a program of research to determine the characteristics and traits required and/or desirable for combat effective attack helicopter pilots. The acronym "ACE" refers to AH-1 Combat Effective pilots who are defined by combat performance criteria other than the traditional criterion of five air-to-air kills.

The construction of profiles derived from the study of combat proven attack pilots and the development of methods to select potential candidates for Attack Helicopter (AH-1) transition training are the objectives of the "ACE" program. The accomplishment of this effort involves the following three interrelated sub tasks which are being conducted concurrently.

Sub Task 1. Survey of Combat Proven Attack Helicopter Pilots to Develop Profiles of Potential "ACE" Pilots and Predictive Instruments to Select Them.

Sub Task 2. Development of Rating Forms for Both School and Unit Level Application to Assess Desirable Traits and Characteristics and Identify Potential Attack Pilots from Among Candidates for Training.

Sub Task 3. Evaluation and Assessment of AH-1 Trainees Against Characteristics and Traits Determined in Research Tasks 1 and 2.

This paper describes the operational procedures and results of the survey of combat proven attack helicopter pilots (Sub Task 1). However, the results and conclusions presented should be interpreted heuristically, as part of a larger research project in progress, which will be finished with the completion of follow up efforts.

There are a number of studies available to indicate that a measurable relationship exists between attitude variables and effectiveness in aerial combat (e.g., Knoell, 1953; Trites, et al, 1953; Strawbridge and Kahn, 1955; Torrence, et al, 1957; Youngling, et al, 1977). Likewise, there is strong

evidence to support the notion that background/biographical material is related to the combat success of aviators (Bond and Burchell, 1944; Torrance, (n. d.); Torrance, et al, 1957; Youngling, et al, 1977).

None of the studies available in the literature has dealt with the relationship between these variables and the combat effectiveness of attack helicopter pilots. However, the working assumption was made that attitude and background variables would also apply to the problem of identifying the population of combat effective attack helicopter pilots.

The method adapted for assessing of the potential value of these variables was to mail questionnaires to samples of "ACE" pilots and controls and then analyze their responses to determine if these variables provide a means for discriminating between individuals in the two groups. The following sections will describe these procedures in detail.

METHOD

SAMPLES.

The Army's Military Personnel Center (MILPERCEN) provided the names and unit addresses of the following two samples of aviators from Officer Master File Records.

ACEs: The "ACE" sample (actually a sub-population) consisting of all commissioned and warrant officer aviators meeting the following criteria:

1. Recipient of the Silver Star or a higher award for valor.

2. Served in Vietnam during the period 1965-1972.

3. Attack Helicopter rated.

A total of 280 officers who met these criteria were included in the "ACE" group. Only aviators on active duty were included.

Controls: The control group consisted of a "random" sample of commissioned and warrant officer aviators meeting the following criteria:

1. Had not received the Distinguished Flying Cross or a higher award for valor.

2. Served in Vietnam during the period 1965-1972.

3. Not gunship qualified.

A total of 385 officers who met these criteria were included in the sample. Only aviators who were still on active duty were included in the study.

SURVEY MATERIALS.

Addressees in the ACE and Control groups all received a set of questionnaires which included:

a. Letter of Instruction

b. Military Background Form (MBF) 15 Items

c. Background and Activities Inventory (BAI) 30 Items

d. Aviator Attitude Questionnaire (AA) 35 Items

e. Self-Description Form (SD) 25 Items

The letter of instruction explained the purpose of the survey in very general terms, assured the respondents that they would remain anonyomous, and requested that the questionnaires be completed and returned to ARI at Fort Rucker. The Military Background Form was designed to get information regarding such variables as the aircraft qualifications, flying experience, combat experience and decorations for valor of the respondents. In addition, it was intended to provide validation of the samples provided by MILPERCEN.

The Background and Activities Inventory contains items selected from existing Army inventories (e.g., the Biographical Inventory of the Flight Aptitude Selection Tests (FAST) (Kaplan, 1965), Interest-Opinion Questionnaire) on the basis of preliminary "hypothesis" obtained from the research literature, structured interviews of Attack pilots, and preliminary item analysis of the responses of more than 50 Attack pilots as compared with student pilots.

The Aviator Attitude Questionnaire was developed from the content analysis of the comments of 10 Attack pilots during unstructured interviews at Ft Rucker. All items in the questionnaire were mentioned by at least two of the pilots interviewed.

The items in the Self-Description Form were drawn primarily from the Self-Description Form of the FAST battery. In addition, the themes for several items were obtained from standardized personality tests (Butcher, 1969).

The Military Background Form and sample items from each of the questionnaires are included in Appendix A.

PROCEDURE.

The packages of survey questionnaires were mailed to ACEs and Controls during a four month period from Dec 76 to March 77. Addressees in both groups received an identical set of questionnaires, except that each questionnaire in a set was stamped with a 3 digit identification number. The two samples were assigned different ID number sequences. Therefore,

the set of questionnaires completed by a single individual all bore the same three digit number. The range within which that number fell was determined by the group the individual belonged in (e.g., 001-449 Controls; 501-999 ACE).

As questionnaires were returned, data was coded for both groups. Survey returns received after mid June were not included i the analyses reported below. Seventy-four percent (74%) of the ACE sample (208) and 58% of the Controls (224) returned completed survey questionnaires. When questionnaires which were returned as undeliverable are taken into account, the return rate was 79% for the "ACE" sample and 65% for the Controls.

RESULTS

Analysis of Questionnaire Items.[1]

Analysis of categorical data of the MBF and the personal data section of the BAI reveal several significant differences between the ACEs and Controls; the responses of the ACE pilots indicate that they differ significantly from Control pilots by being: (1) of higher rank ($p = .0002$), (2) Have more time in service, (3) more likely to be in the combat arms ($p < .0001$), (4) have more combat experience ($p < .0001$), (5) more likely to report having higher efficiency ratings ($p = .033$), and (6) lower in civil education level ($p = .027$) (see Table 1 below). Although the ACE group were of higher rank (see Table 1) and had more median years of service than the Non ACE Controls (14 yrs vs 12 yrs, $p < .01$), they did not differ significantly from the Controls in median age (ACE = 35 yrs, Controls = 34 yrs). The combat arms affiliation and combat experience variables are predictable, and do confirm the effectiveness of the sampling criteria in identifying combat experienced attack pilots. The principal source of the difference between ACE and Control pilots in civil education is due to the smaller percentage of ACEs (35%) than Controls (51%) who have a college degree. The questionnaire items for which the responses of the ACEs and Controls differ significantly are presented in Table 1. The BAI items, other than the civil education, which were significant indicate that, as a group, the ACEs differ from the Controls by reporting that they engage and excel more often in sports and activities of a physical or active nature.

The results also indicate that differences in interest items which are not of a physical or active nature do not discriminate between the combat effectives and Controls. None of the items asking about the frequency of engaging in activities such as reading, hobbies, special interests, etc.

[1] It was possible to include data for 208 ACEs and 222 Non ACE Controls in the analysis. However, because of missing data the Ns for a specific item were often slightly smaller.

TABLE 1 STATISTICALLY SIGNIFICANT QUESTIONNAIRE ITEMS

MILITARY BACKGROUND FORM

MBF 3

GRADE	N	W1	W2	W3	W4	1LT	CPT	MAJ	LTCOL	COL	χ^2 [1]	df	p.
		10	11	12	13	15	16	17	18	19	30.303	8	.0002
ACE	205	0	6.3	3.9	1.0	.5	38.0	28.8	16.1	5.4			
CONTROL	221	.5	1.8	.9	4.1	0	48.9	29.0	14.9	0			

MBF 10

COMBAT EXPERIENCE

	N	None	1-6	7-12	13-24	More 24	χ^2	df	p.
ACE	205	0.0	0	18.0	63.0	19.0	45.471	4	<.0001
CONTROL	221	.9	4.5	42.5	43.4	8.6			

MBF 15

Where do your efficiency ratings fall in relation to your peers?

	UPPER THIRD	MIDDLE THIRD	LOWER THIRD	DON'T KNOW	χ^2	df.	p.	N
ACE	73.9	13.8	2.5	9.9	8.761	3	.033	203
CONTROL	61.8	17.3	2.3	18.6				220

[1] The phi and Cramer's V coefficients for significant items fell in the .10 to .30 range. They can be computed exactly from the values in the table.

BACKGROUND ACTIVITIES INVENTORY

BAI 4

How far did you go in school before you came into the Army (or other branch of the Armed Forces) on extended active duty?

	ACE	N	CONTROL	N	χ^2	df.	p.
A. Less than high school	5.3	207	2.3	222	15.76760	7	.0273
B. High school equivalent (GED or other equivalency)	2.9		.9				
C. High school graduate	21.7		16.7				
D. College, less than two years	15.5		11.7				
E. College, two years or more	16.4		15.3				
F. College degree	35.3		50.5				
G. Graduate work, no degree	2.9		1.8				
H. Graduate or professional degree	0		.9				

Mark how often you have participated in the activity listed according to the following scale:

A = Never

B = Once or Twice

C = A Number of Times

D = Frequently

BAI 10	N	A	B	C	D	χ^2	df.	p.
Water Skiing						8.713	3	.0334
ACE	206	5.8	16.0	36.9	41.3			
CONTROL	222	10.4	22.5	36.9	30.2			
BAI 11								
Boxing or Wrestling								
ACE	208	9.6	35.1	38.0	17.3	8.419	3	.0381
CONTROL	222	14.0	36.9	40.5	8.6			
BAI 12								
Drag Racing						7.964	3	.0468
ACE	207	37.7	33.3	21.7	7.2			
CONTROL	222	49.5	31.1	15.3	4.1			
BAI 15								
Fist Fights						15.763	3	.0013
ACE	208	13.5	55.3	26.9	4.3			
CONTROL	222	23.0	59.5	16.7	.9			

AVIATOR ATTITUDE

CHECK: T For True

F For False

	GROUP	N	T	F	CHI SQ	df.	p.
AA 2							
A guy may be a turkey when it comes to aircraft control, but if he's an aggressive competitor he'll be a good combat pilot.	ACE	208	17.3	82.7	5.73106	1	.0167
	CONTROL	221	9.0	91.0			
AA 4							
If I have to be in a combat assignment, I want to be doing the shooting.	ACE	207	90.3	9.7	80.05700	1	<.0001
	CONTROL	220	50.0	50.0			
AA 7							
In case of another war, I want a combat aviation assignment.	ACE	208	86.1	13.9	27.49691	1	<.0001
	CONTROL	222	63.5	36.5			
AA 8							
I hate to admit that I watch TV most of the time when I'm off duty.	ACE	205	9.8	90.2	4.99292	1	.0255
	CONTROL	220	17.7	32.3			

AVIATOR ATTITUDE

CHECK: T For True
F For False

	GRADE	N	T	F	CHI SQ	df.	p.
AA 9 Without good attack helicopter teams, the US Army won't have much of a chance in the next war.	ACE	207	66.2	33.8	4.53620	1	.0332
	CONTROL	221	55.7	44.3			
AA 11 The peacetime Army doesn't offer me enough challenge.	ACE	207	26.6	73.4	6.26552	1	.0123
	CONTROL	222	16.2	83.8			
AA 14 The best aviators are highly competitive in all they do.	ACE	208	73.1	26.9	3.71334	1	.0540
	CONTROL	222	64.0	36.0			
AA 20 The best aviators are egotistical elitists.	ACE	207	16.9	83.1	4.60088	1	.0320
	CONTROL	222	9.5	90.5			

AVIATOR ATTITUDE

CHECK: T For True

F For False

Item	GRADE	N	T	F	CHI SQ	df.	p.
AA 21 I may be biased, but Army aviation will be where the real action is in the next war.	ACE	206	74.3	25.7	15.79341	1	<.0001
	CONTROL	222	55.4	44.6			
AA 25 I enjoy the power of weapons.	ACE	207	65.7	34.3	28.20239	1	<.0001
	CONTROL	220	39.5	60.5			
AA 27 Give me the agressive, active pilot for the gunship assignments every time.	ACE	204	67.2	32.8	4.69071	1	.0303
	CONTROL	218	56.4	43.6			
AA 28 Debating may be alright sometimes, but give me a good verbal free-for-all where you can really mix things up.	ACE	205	44.9	55.1	14.57927	1	<.0001
	CONTROL	221	26.7	73.3			

AVIATOR ATTITUDE

CHECK: T For True
F For False

Item	GRADE	N	T	F	CHI SQ	df.	p.
AA 29 Aviation is the main topic of discussion among my friends during off duty hours.	ACE	206	26.7	73.3	9.07395	1	.0026
	CONTROL	221	14.5	85.5			
AA 30 In a real tight situation, I act first and explain later.	ACE	205	89.3	10.7	5.70098	1	.0170
	CONTROL	220	80.5	19.5			
AA 33 If it involves a combat assignment, Army aviation is not for me.	ACE	207	1.0	99.0	4.55676	1	.0328
	CONTROL	221	5.0	95.0			
AA 34 Exceptionally good eyes and hands are more important than aggressiveness to a gunship pilot.	ACE	206	35.9	64.1	12.76887	1	.0004
	CONTROL	218	53.7	46.3			

SELF DESCRIPTION FORM

CHECK: Y For Items That Describe You
N For Items That Do Not Describe You

	GROUP	N	Y	N	CHI SQ	df.	p.
SD 2							
I don't like to argue.	ACE	208	42.3	57.7	6.85661	1	.0088
	CONTROL	222	55.4	44.6			
SD 3							
I believe in getting even.	ACE	208	33.7	66.3	5.54166	1	.0186
	CONTROL	222	23.0	77.0			
SD 6							
I get mad easily and get over it quickly.	ACE	208	42.3	57.7	3.67998	1	.0551
	CONTROL	222	32.9	67.1			
SD 8							
I don't like doing things on the spur of the moment.	ACE	207	22.2	77.8	6.02391	1	.0141
	CONTROL	222	33.3	66.7			

ment of a quad track for cargo/lift pilots have also been discussed. These developments will have the following effects on a differential selection program for Army aviators: (1) emphasis must shift to selection of flight training students rather than experienced aviators, (2) a multiple classification system based on the results of discriminant analysis of effective aviator specialists must be developed. This means that questionnaire items to be effective with IERW students will have to be general in nature, i.e., they cannot be tailored to the opinions and attitudes of experienced Army aviators as are many of the Aviation Attitude items. The multiplication of tracks will probably result in the identification of a second discriminant function reflecting a set of technical specialization items to supplement the combat oriented discriminant function identifiable in this research. It is anticipated that the centroids of the aviation specialities which would result from a multi-track system would be meaningfully represented in such a discriminant space.

SUMMARY

The responses of 208 combat effective attack helicopter pilots (ACEs) and 222 control pilots to four questionnaires, (1) Military Background Form, (2) Background and Activities Inventory, (3) Aviator Attitude Questionnaire and (4) Self Description Form, were analyzed. Items on which the two groups differed significantly were identified. It is anticipated that the items will provide the content for the construction of profiles of individuals who are suited for training as attack pilots. Discriminant analysis of the data indicates that a single, shorter questionnaire can be developed to classify pilots as potential AH-1 combat effectives(ACEs) or non-ACEs.

produced any significant differences between the two groups. The Aviator Attitude items that were the most discriminating, as noted above, were derived from the analysis of comments by combat experienced attack pilots obtained during interviews. Because of this, much of their effectiveness may be limited to selection of experienced aviators for advanced training. However, some of the more significant items may be of a general nature, (e.g., "I enjoy the power of weapons"). The differences between the ACEs and Controls on these items are consistent with stereotyped notions of how the aggressive combat performer "should" differ from his less aggressive peers.

Discriminant Analysis.

Discriminant analysis was applied to the questionnaire data to identify the combination of items that best differentitate between the "ACE" and Control helicopter pilots. The stepwise discriminant analysis procedure was used (MINRESID method) to eliminate less useful items from the discriminant functions (Nie, et al, 1975). An arbitrary maximum number of 15 steps were specified for the analysis. This was done to obtain a reduced set of items for inclusion in a single new instrument along with additional "filler" and untried items. The summary table for the analysis is shown in Table 2. As indicated in the table by the 15th step the

TABLE 2

SUMMARY TABLE OF DISCRIMINANT ANALYSIS OF QUESTIONNAIRE DATA

Step	Item[1] Entered	F to Enter Or Remove	Change In Rao's V	Significance
1	AA 04	94.26	94.23	< .001
2	AA 07	14.22	17.98	< .001
3	SD 13	9.76	12.86	< .001
4	BAI 16	10.97	14.89	< .001
5	AA 34	8.06	11.31	< .001
6	AA 01	7.41	10.66	< .001
7	SD 02	6.21	9.13	< .003
8	BAI 14	6.43	9.65	< .002
9	BAI 04	4.80	7.36	< .007
10	AA 16	6.32	9.84	< .002
11	AA 02	4.87	7.75	< .005
12	BAI 27	3.82	6.18	< .013
13	SD 10	3.13	5.12	< .024
14	SD 08	2.77	4.60	< .032
15	AA 24	2.51	4.21	< .040

[1]See Appendix A for content of items.

contribution of any additional variables to the discrimination was approaching insignificance (change in Rao's V, p =.04). The items selected for inclusion in the discriminant function were for the most part included in the set of statistically significant items in Table 1. In addition, several apparent suppressor items were selected. The complete set of 15 items are presented in Appendix A. The set of original variables included in the discriminant function were then used to classify members of both the original samples to see how many would be correctly classified into their actual group. By this process the probability of belonging to one or the other groups is calculated by separate linear combination of the variables for each individual case, then assignment is made to the group with the highest probability. As is shown in Table 3, this procedure resulted in a 80% probability of correct classification of the ACE and Control pilots in their actual groups.

TABLE 3 CLASSIFICATION OF GROUP MEMBERSHIP

Actual Group	Predicted Group	
	ACE	Control
ACE	84.7%	15.3%
CONTROLS	24.5%	75.5%

A total of 69.9% of cases correctly classified

This relatively high % of correct classification is a necessary prerequisite for using the discriminant function to predict or classify a new or unknown sample of aviators into the group their responses indicate they are most suited for.

DISCUSSION AND CONCLUSIONS

The analysis of the questionnaire data indicates that personality, attitude and background measures can be used to discriminate between combat effective attack helicopter pilots and control pilots. It is apparent that further research to develop instruments to select helicopter pilots for attack training is worth pursuing. It is anticipated that, in addition to providing training selection/classification instruments, the content of the items can provide the basis for developing representative profiles to assist unit commanders in assigning aviators for training as attack pilots.

The BAI items, for example, suggest that ACE pilots are more likely to report that they participate, and are proficient, in physical/aggressive

activities, Target shooting, combatives (boxing, wrestling) and strength/endurance/fitness type activities are all more typical of the combat effective. This suggests that the ACE pilots are individuals who have an image of themselves as highly physically active, aggressive and are confident in their ability to perform activities which conform to this image (whether objectively justified or not).

The SD items for which there was a significant difference between ACE and Control pilots indicate that the combat effectives are more likely to be active, argumentative, patriotic individuals who like taking risks, but want to feel they are in control of their actions. They are also more quick to anger and be vindictive than the Non ACE Controls.

The AA items, as mentioned above, are aimed at tapping the attitudes of experienced Army aviators and to that extent are limited in reflecting more general attitudes. However, the following generalizations are possible: The "ACE" pilots as a group are more likely (1) to express a desire for combat duties, (2) enjoy power of weapons, (3) feel that their job is important or critical, (4) to feel that aggressiveness is important for combat, and (5) enjoy heated arguments.

The adequacy of the discriminant function in classifying the original set of cases was relatively high (80%). This indicates that the discriminant function derived from the analysis of ACE and Non ACE Control pilots can be used to classify aviators of unknown group membership. This will be accomplished by calculating two classification scores, i.e., ACE and Non ACE, for each aviator and classifying him into the group with the highest score. In this way questionnaires can be used in conjunction with other information and data (e.g., ratings), to assist commanders in assigning aviators to types advanced training (e.g., Attack AH-1 vs Cargo/lift CH-47).

Research is currently underway to determine the predictive validity of a rating form for AH-1 pilot candidates (Eastman and McMullen, 1976). As an indirect result of this research, the training performance grades of a large sample of AH-1 pilots will be available. Follow-up research is planned to try and classify high scoring AH-1 graduates and a group of non AH-1 rated pilots using the classification functions developed from the discriminant analysis of these surveys. In addition, new items will be added to assess areas not included in the present questionnaires. These will include the following: (1) question the aviators about childhood experiences regarding risk taking, aggressive behavior and troublemaking behavior, (2) measures of confidence in physical prowess, and resistance to stress and (3) ask whether he "sought out" or "drifted into" a military career.

The research objectives of this program were to develop instruments for selecting aviators for advanced training. However, a development in the Army's Initial Entry Rotary Wing (IERW) program has been the initiation (in June 76) of a dual track program in which 25% of the students will be trained in the OH-58 Scout for the final (tactics) phase of IERW; A triple track program in which another sizable percentage of IERW students will be tracked into an AH-1 attack helicopter transition and tactical training as an attack pilot has been scheduled for 1978. Plans for eventual establish-

ment of a quad track for cargo/lift pilots have also been discussed. These developments will have the following effects on a differential selection program for Army aviators: (1) emphasis must shift to selection of flight training students rather than experienced aviators, (2) a multiple classification system based on the results of discriminant analysis of effective aviator specialists must be developed. This means that questionnaire items to be effective with IERW students will have to be general in nature, i.e., they cannot be tailored to the opinions and attitudes of experienced Army aviators as are many of the Aviation Attitude items. The multiplication of tracks will probably result in the identification of a second discriminant function reflecting a set of technical specialization items to supplement the combat oriented discriminant function identifiable in this research. It is anticipated that the centroids of the aviation specialities which would result from a multi-track system would be meaningfully represented in such a discriminant space.

SUMMARY

The responses of 208 combat effective attack helicopter pilots (ACEs) and 222 control pilots to four questionnaires, (1) Military Background Form, (2) Background and Activities Inventory, (3) Aviator Attitude Questionnaire and (4) Self Description Form, were analyzed. Items on which the two groups differed significantly were identified. It is anticipated that the items will provide the content for the construction of profiles of individuals who are suited for training as attack pilots. Discriminant analysis of the data indicates that a single, shorter questionnaire can be developed to classify pilots as potential AH-1 combat effectives(ACEs) or non-ACEs.

REFERENCES

Bond (Major) and Burchell (Major) A study of 100 successful airmen with particular respect to their motivation and resistance to combat stress. Report of Project #22, submitted to: Surgeon, HQ, Eight Air Force, December 1944.

Butcher, J. N. (ed.) MMPI; Research Development and Clinical Applications. New York: McGraw Hill, 1969.

Eastman, R. F. and McMullen, R. L. Reliability of associate ratings of performance potential by Army aviators. Presented at the 18th Conference of the Military Testing Association, 18-22 October, 1975 (available as ARI Research Memorandum 76-28, Nov 1976.)

Kaplan, Harry. Prediction of success in Army aviation training. ARI Technical Research Report 1142, June 1965.

Knoell, Dorothy M. Relationships between attitudes of bomber crews in training and their attitudes and performance in combat (AFPTRC-TN-56-49). Lackland Air Force Base, Texas; Air Force Personnel and Training Research Center, April 1956.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K. and Bent, D. H. SPSS Statistical Package for the Social Sciences, 2nd Ed., New York: McGraw Hill, 1975.

Strawbridge and Kahn. Fighter pilot performance in Korea. Chicago, Illinois Institute for Air Weapons Research, November 1955 (IAWR Report #55-10).

Torrance, E. P. The development of a preliminary life experience inventory for the study of fighter effectiveness. Lackland Air Force Base, Texas.

Torrance, E. P., Rush, C. H., Kohn, H. B., Doughty, J. M. (AF PTRC-TR-57-11). Factors in fighter interceptor pilot combat effectiveness, 1957, NTIS No. AD-146-407.

Trites, D. K., Holtrman, W. H., Templeton, R. C., Sells, S. B. Psychiatric Screening of Flying Personnel: Research on the SAM Sentence Completion Test. (AF Project No. 21-0202-0007, #3), USAF School of Aviation Medicine, Randolph Field, Texas, July 1953.

Youngling, E. W., Levine, S. H., Mochamuk, J. B. and Weston, L. M. Feasibility Study to predict combat effectiveness for selected military roles: fighter pilot effectiveness. MOC E1634, 29 April 1977.

APPENDIX A

ITEMS SELECTED FOR DISCRIMINANT FUNCTION

ITEM ID	(Check)	ITEM CONTENT
AA04	T = True F = False	If I have to be in a combat assignment I want to be doing the shooting.
AA07	T = True F = False	In case of another war I want a combat aviation assignment.
SD 13	Y = Describe you N = Does not describe you	My behavior is largely controlled by the customs of my society.
BAI 16	How often have you participated? A = Never B = Once or twice C = A number of times D = Frequently	Sport parachuting or sky diving
AA 34	T = True F = False	Exceptionally good eyes and hands are more important than aggressiveness to a gunship pilot.
AA 01	T = True F = False	Some people are naturally adapted to combat.
SD 02	Y = Describes you N = Does not describe you	I don't like to argue.

BAI 14	How often have you participated? A. Never B. Once or Twice C. A number of times D. Frequently	Surfboard riding
BAI 04		How far did you go in school before you came into the Army (or other branch of armed forces) on extended active duty? A. Less than high school B. High school graduate C. High school equivalent (GED or other equivalent). D. College, less than two years E. College, two years or more F. College degree G. Graduate work, no degree H. Graduate or professional degree
Al 16	T = True F = False	The best scout and gunship pilots have strong suicidal tendencies.
AA 02	T = True F = False	A guy may be a turkey when it comes to aircraft control, but if he's an aggressive competitor he'll be a good combat pilot.
BAI 27	How well do you perform? A. Outstanding B. Well C. Adequately D. Poorly E. Do not engage actively	Weight lifting or strength exercises.

SD 10	Y = Describes you N = Does not describe you	I sometimes tease animals.
SD 08	Y = Describes you N = Does not describe you	I don't like doing things on the spur of the moment.
AA 24	T = True F = False	I'm among the best but the Army isn't giving me any credit for it.

302

MILITARY BACKGROUND FORM

DATE: ______________________ AGE: ______________________

CURRENT DUTY STATION: ______________________

GRADE: ______________________ BRANCH: ______________________

	HS	YRS COLLEGE COMPL 1	2	3	4	GRADUATE STUDY	DEGREES AWD (IF APPL)
HIGHEST CIVILIAN EDUCATION LEVEL ATTAINED:	( )	( )	( )	( )	( )	( )	________

YEARS OF MILITARY SERVICE: ____________ YEAR GRADUATED FROM FLIGHT SCHOOL: ____________

		UH-1	AH-1	OH-58	CH-47		YES	NO
AIRCRAFT QUALIFICATION:	R/W	( )	( )	( )	( )	F/W	( )	( )

	None	1-6 mos	7-12 mos	13-24 mos	More than 24 mos
COMBAT EXPERIENCE:	( )	( )	( )	( )	( )
AVIATION COMBAT ASSAULT EXPERIENCE:	( )	( )	( )	( )	( )
COMBAT EXPERIENCE AS A GUNSHIP/ATTACK PILOT:	( )	( )	( )	( )	( )

DECORATIONS FOR VALOR (IF APPL): ______________________

____

DECORATIONS AWARDED FOR VALOR DURING AVIATION MISSIONS (IF APPL): ________

____

	UPPER THIRD	MIDDLE THIRD	LOWER THIRD	DON'T KNOW
WHERE DO YOUR EFFICIENCY RATINGS FALL IN RELATION TO YOUR PEERS?	( )	( )	( )	( )

Pilot Selection Research in the Air Force

David R. Hunter
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

As part of the mission of the Personnel Research Division of the Air Force Human Resources Laboratory, continuing studies have been conducted to investigate ways to improve the selection procedures for admission to Undergraduate Pilot Training (UPT). This research has included attempts to improve the existing paper-and-pencil selection measures, investigations into the use of new solid-state psychomotor apparatus tests, and evaluations of learning ability through the use of a light-plane simulator. This report will outline the research that has been performed by the Personnel Research Division in these areas and will indicate some of the tasks that remain to be addressed.

Since the early 1950's, when psychomotor testing was discontinued, the principal selection instrument for pilot training has been the Pilot Composite of the Air Force Officer Qualifying Test (AFOQT). This composite, in the previous (Form M) version of the AFOQT, consisted of the seven subscales shown in Table 1. Also shown in Table 1 are the eight subscales which comprise the pilot composite in the version (Form N) of the AFOQT which will shortly become operational.

As can be seen from comparing these two lists, there is a considerable difference between the two forms. This modification arose out of a search for means to improve the predictive validity of the Pilot Composite through the use of different item types.

An experimental reference battery consisting of the 21 scales listed in Table 2 was administered to a sample of officers and office trainees slated to attend UPT and validated against their performance. The correlations for each of these scales with a dichotomous Pass/Fail criterion and a dichotomous Pass/Flying Training Deficiency (FTD) elimination criterion are also presented in Table 2. Based upon the results of this study, five new scales were selected for the Form N AFOQT.

Additionally, as a result of research performed by Guinn, Vitola, and Leisey (1976), a Biographical and Attitude Scale was included. This scale was developed as a result of research using the Strong Vocational Interest Blank (SVID) and the Officer Background and Attitude Survey. The correlations between measures developed by Guinn et al. from these two instruments and the two training criteria are shown in Table 3. In contrast to most of the measures developed for screening personnel entering UPT, these scales seem to have greater validity for prediction of overall attrition than for Flying Training Deficiency elimination, thus possibly indicating that they do not tap the abilities related to flying skill but rather the attitudes and habits that allow one to succeed.

Table 4 presents the multiple correlations between the paper-and-pencil measures and the Pass/Fail UPT criterion. As can be seen from this table, the multiple correlations are typically of a rather low order and, in most cases, did not achieve statistical significance. Only the SVIB consistently makes a significant contribution to prediction of the

criterion; however, the multiple correlation reported for the AFOQT is seriously attenuated by restriction of range due to preselection using this test.

To examine the relative contributions of these measures, the null hypotheses listed at the bottom of Table 4 were tested with the noted results.

In general, it seems that it would be possible to obtain significant increases in the predictive validity of the AFOQT through the addition of a scale measuring attitudes and interests, and this has been done for the latest revision of the AFOQT.

The second area of pilot selection research has been concerned with the measurement of psychomotor abilities and their relation to success in UPT. Sanders, Valentine, and McGrevy (1971) and McGrevy and Valentine (1974) have reported on the development and validation of two aircrew psychomotor tests which have shown promise as possible instruments for pilot selection.

The first of these tests, Two-Hand Coordination, requires that the subject track a moving target with a small X shaped cursor, using two hand joysticks. The right-hand joystick controls the movement of the cursor in the right-left (X) coordinate, while the left-hand joystick controls the movement of the cursor in the up-down (Y) coordinate. Figure 1 shows the display used in this test and in the second test--Complex Coordination.

The Complex Coordination test involves a compensatory tracking task in which the subject controls the X and Y displacements of a cursor with

a single, large floor-mounted joystick. The subject's task is to keep the cursor as close as possible to the intersection of a vertical and horizontal line of dots. At the same time, he must use a rudder bar to keep a short bar of light aligned with the vertical row of dots.

Scores for both of these tests consist of the summed absolute displacements (in CRT units, approximately .01 inch) from the target point to the cursor in the X and Y axes and the right-left axis (Z-axis) for the rudder bar. These displacements are summed over the five 1-minute periods of the tests for the X, Y, and Z axes separately.

Table 5 presents the correlations of these measures with UPT criteria for two independent samples. The first sample (from an unpublished study by McGrevy & Valentine, 1975) consisted primarily of officer trainees slated to attend UPT. The correlations between scores from the fourth and fifth minutes of the two tests for the X, Y, and Z axes with UPT Pass/Fail and with Flying Training Deficiency Elimination versus any other disposition.

The second sample consisted principally of officers about to attend the Flight Screening Program (FSP) at Hondo, Texas. The FSP program is the first phase of UPT and consists of about 15 hours of instruction in a T-41 aircraft.

Correlations are reported here for the arithmetic sum of the scores from minutes four and five for the two tests. As can be seen from this table, these two tests correlate significantly and consistently with performance in UPT.

As a result of the initial studies conducted using these two tests, which used a small minocomputer for the generation and scoring of the tests, two new portable test devices were obtained. One of these devices is shown in Figure 2. These devices are entirely self-contained and use solid state electronic components to increase reliability and decrease problems of calibration which contributed to the discontinuance of psychomotor testing for the selection of pilot trainees in the early 1950's. A continuing research program is under way to assess the validity and reliability of these devices, with the expectation that they may be included in the screening process for the selection of pilot trainees.

The third major area of pilot selection research has focused on the evaluation of learning ability using an experimental task highly similar to that involved in pilot training. An Automated Pilot Aptitude Measurement System (GAT) was devised by Long and Varney (1975), which utilized two light-plane simulators (Link-Singer General Aviation Trainers) interfaced to a minicomputer to collect performance data on a number of flight parameters during the course of a 5-hour syllabus of instruction on how to fly the simulators.

Two studies were conducted using this system. The first study was reported by Long and Varney (1975), and data from that study has been reanalyzed so as to ensure comparability with the data analysis performed on the second study. Because of the large number of variables (190) collected on each subject by the system, some method of data reduction was considered desirable. The Long and Varney study used a factor

analytic approach to reduce the number of variables; therefore, that approach was replicated in the reanalysis of the data.

A principal components factor analysis, followed by Varimax rotation, was performed on the 190 variables obtained from the system. Figure 3 shows the eigenvalues for the first ten unrotated factors. Figure 4 shows the proportion of variance for the first ten rotated factors. From an examination of these two figures, it would seem that further examination should be limited to the first six rotated factors. Based upon their loadings with the raw variables, these first six factors may be identified as shown in Table 6.

The correlations of these six factors with the two UPT criteria are shown in Tables 7 and 8 for Study I and Study II, respectively. There is considerable variation in the validities from one study to the next, and this can be attributed, at least in part, to the instability of factors determined from 190 variables while using only 140 subjects.

Because of this instability, it was considered desirable to develop a simpler, more stable data reduction procedure. The simplest and most obvious procedure was, of course, to simply form sums or averages of the sets of the 190 variables that were in the same metric--that is, to simply take the average of all deviations from command heading, all deviations from command altitude, and so on. Tables 9 and 10 show the correlations between these simple CAT scores and the two training criteria for the two studies.

The correlations here are as high, or higher, than the ones reported for the CAT factors, while retaining their validity across the two studies.

The multiple correlations of these five simple scores with the two criteria are given in Table 11 for Study I. Table 12 shows the multiple correlations of the six factor scores from Study I and also the cross-validated multiple correlations obtained from Study II. These results show that the GAT factors did not retain their validity for the prediction of the Pass/Fail criterion, and there was a substantial decrease in validity for the prediction of the Pass/FTD criterion. As noted earlier, these decreases may be attributed, in part, to the instability of the factors based upon such a relatively small sample. While the cross-validated multiple correlations for the simple GAT scores are not yet available, it is expected that there will be considerably less shrinkage in these correlations. Table 13 gives an approximation of these shrunken multiple correlations, calculated using the standard correction for shrinkage.

In Table 14, the multiple correlations are reported for some of the combinations of variables used in Study I. In general, it seems that the best combination to achieve maximum validity should involve the psychomotor and GAT variables. Although the GAT factor scores are used in these regression problems, the same order of validity should be achieved through using the simple GAT variables, with considerably less shrinkage.

While the GAT variables can make significant contributions to UPT prediction, analyses performed by AFHRL staff members have indicated that the use of a selection system which included the GAT along with paper-and-pencil and psychomotor measures, may prove economically unfeasable under the present pilot flow conditions, as compared with a selection system

which consists of only the paper-and-pencil and psychomotor measures. This is due not so much to the cost of the testing apparatus, but rather to the cost of transporting applicants to the testing sites and providing for them while they are undergoing testing. Because of these considerations, further research is contemplated which will examine the degree to which validity comparable to that obtained with the GAT may be obtained from smaller, more portable, low fidelity devices. Specifically, whether a "desk-top" light-plane simulator, without motion base, or possibly a simple cathode ray tube display and small minicomputer might be used to replace the GAT system while measuring the same critical learning abilities.

In conclusion, then, it has been found that the validity of paper-and-pencil measures may be increased through the addition of items which address the background and attitudes of the applicants. Furthermore, the use of psychomotor measures can make a significant contribution through the measurement of abilities not tapped by paper-and-pencil measures.

The use of learning sample measures, such as those obtained from the GAT system, can also make significant independent contributions to increased validity; however, additional research must be performed to develop a more cost effective vehicle for this type of testing.

REFERENCES

Guinn, N., Vitola, B.M., & Leisey, S.A. Background and interest measures as predictors of success in undergraduate pilot training (AFHRL-TR-76-9). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, May 1976.

Long, G.E., & Varney, N.C. Automated pilot aptitude measurement system (AFHRL-TR-75-58). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, September 1975.

McGrevy, D.F., & Valentine, L.D. Validation of two aircrew psychomotor tests (AFHRL-TR-74-4). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, January 1974.

Sanders, J.H., Valentine, L.D., & McGrevy, D.F. The development of equipment for psychomotor assessment (AFHRL-TR-71-40). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, July 1971.

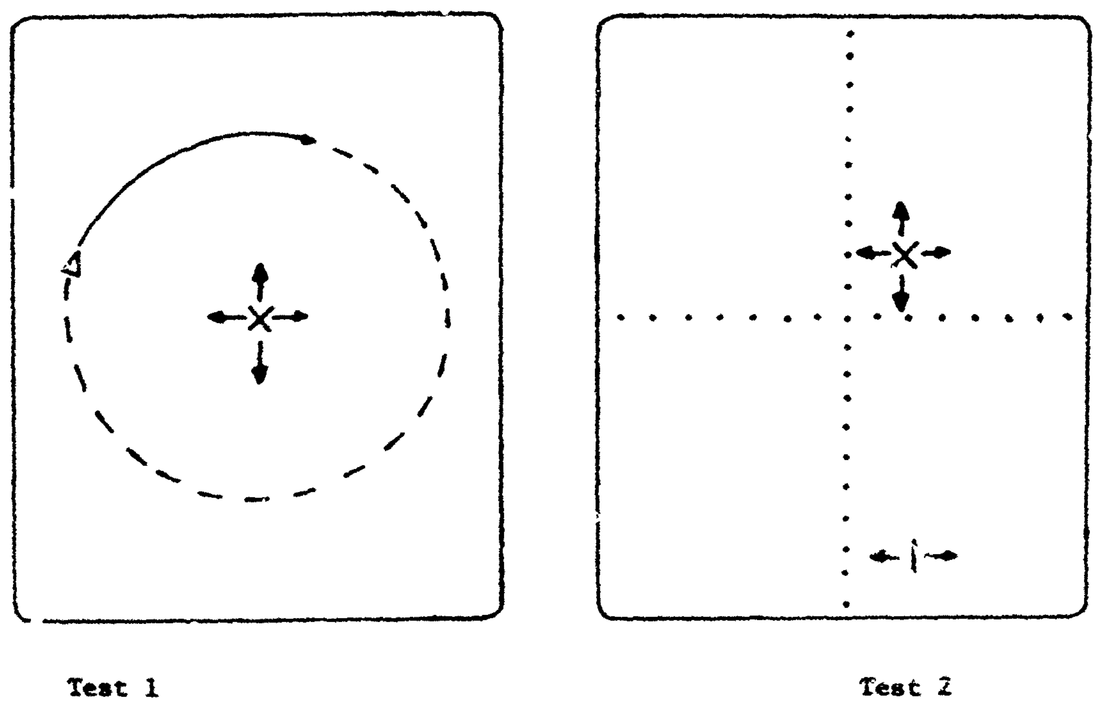


Figure 1. Psychomotor Displays

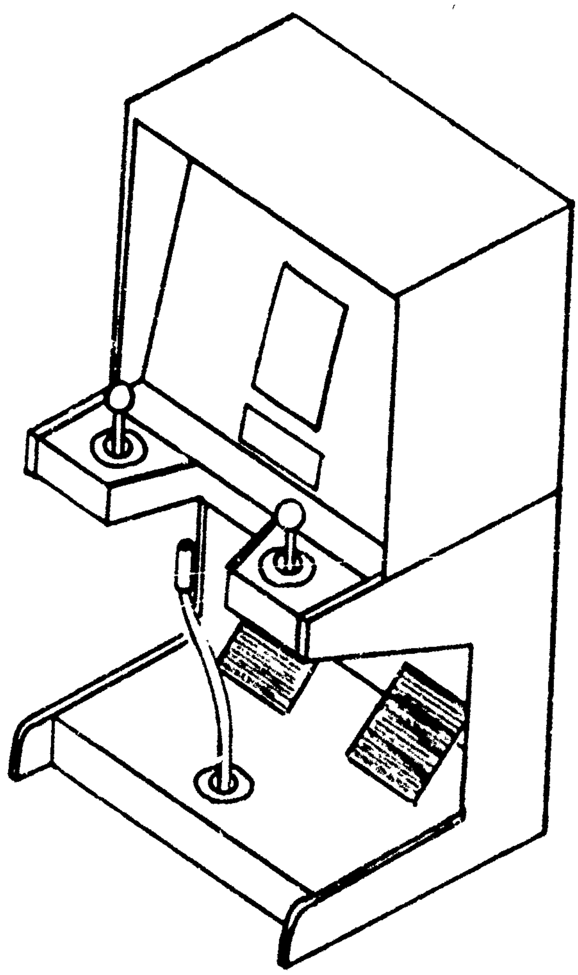


Figure 2. Portable Psychomotor Device

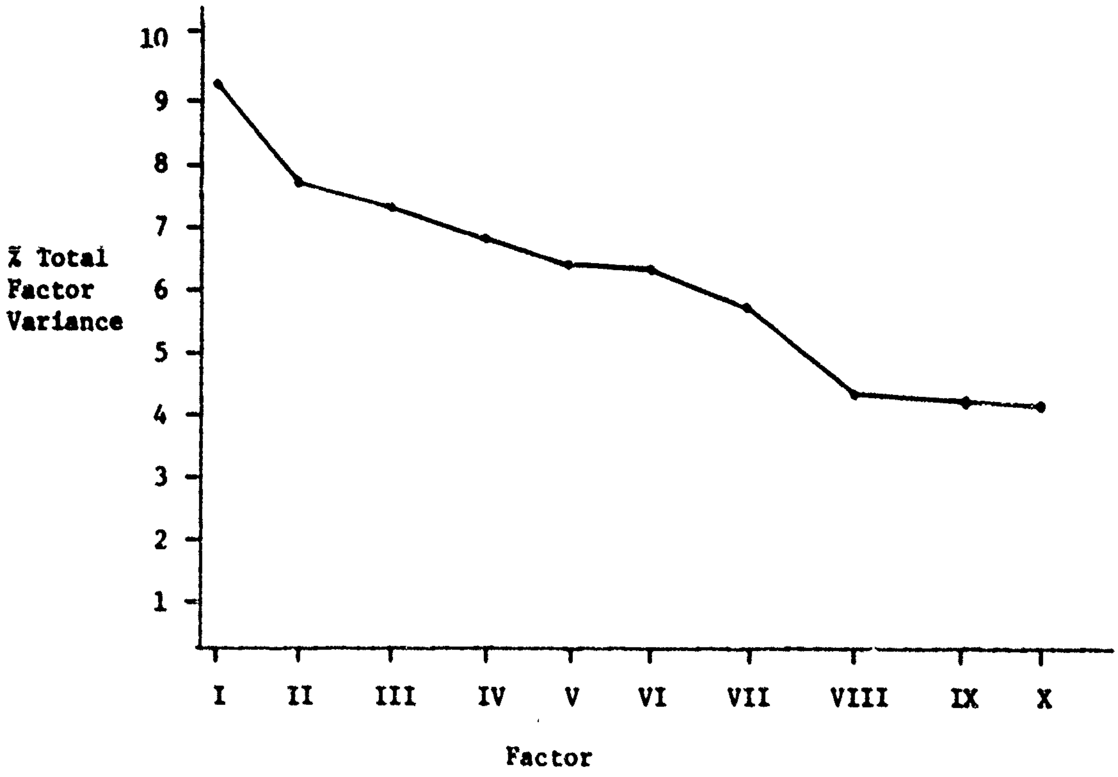


Figure 3. Proportion of Variance by Rotated GAT Factors - Study I

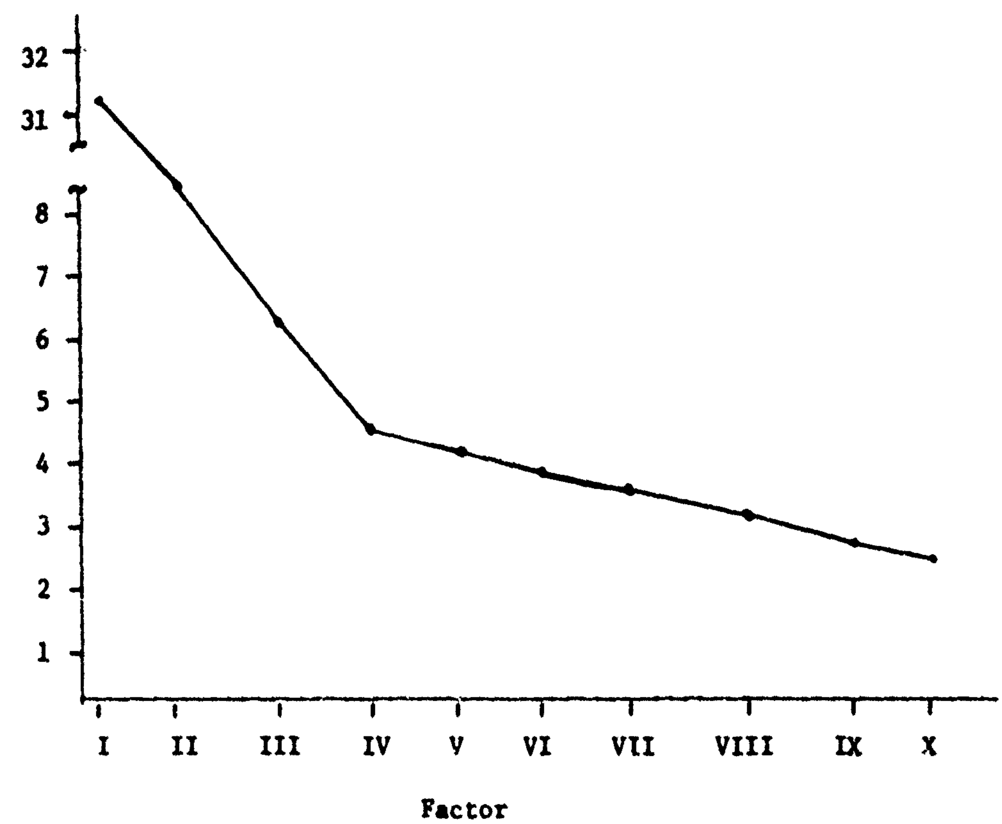


Figure 4. Eigenvalues - Factored GAT Matrix

Table 1. Pilot Composite Subscales

Form M	Form N
Mechanical Information	Verbal Analogies
Mechanical Principles	Table Reading
Pilot Biographical Inventory	Electrical Maze
Aviation Information	Block Counting
Visualization of Maneuvers	Scale Reading
Instrument Comprehension	Mechanical Comprehension
Stick and Rudder Orientation	Instrument Comprehension
	Pilot Biographical and Attitude Scale

Table 2. Experimental Paper-and-Pencil Scales
(N = 245)

Scale	Pass/Fail	FTD/Other
Scale Reading	.19	-.16
Letter Sets	.10	-.15
Tool Functions	.04	-.08
Electrical Information	.02	-.10
Mechanical Principles	.10	-.12
Word Knowledge	.03	-.15
Word Grouping	-.01	-.04
Verbal Analogies	.13	-.19
Block Counting	.18	-.15
Point Distance	.04	-.06
Electrical Maze	.13	-.14
Pattern Detail	.07	-.14
Rotated Blocks	.08	-.10
Tools	.04	.02
Figure Analogies	-.01	-.04
Hidden Figures	.05	-.03
Answer Sheet Marking	.05	-.09
Table Reading	.17	-.08
Large Tapping	.05	-.10
Trace Tapping	.05	-.03
Discrimination-Reaction	.06	-.02

Table 3. Attitude and Interest Measures

	Pass/Fail (N = 265)	Pass/FTD (N = 227)
<u>SVIB</u>		
Key A	.13*	.06
Key B	.16*	.09
Key C	-.06	-.01

	Pass/Fail (N = 257)	Pass/FTD (N = 220)
<u>OBAS</u>		
Total Examination Scale	.15*	.03
Flying Deficiency Scale	.13*	.03

*$p < .05$

Table 4. Regression Problems Using Paper-and-Pencil Measures for UPT Pass/Fail Criterion

Problem Number	Predictors	N	No. of Predictors	R
1	AFOQT	715	5	.12
2	Reference Battery	745	21	.20
3	OBAS	257	2	.15
4	SVIB	265	3	.18*
5	OBAS & AFOQT	206	7	.23
6	OBAS & Reference Battery	250	23	.25
7	SVIB & AFOQT	214	8	.24
8	SVIB & Reference Battery	258	24	.28
9	SVIB & OBAS	256	5	.23*

Null Hypotheses	Field Model	Restricted	df_1	df_2	F Ratio
OBAS makes no contribution to AFOQT	5	1	2	198	4.02*
SVIB makes no contribution to AFOQT	7	1	3	205	3.13*
OBAS makes no contribution to Reference Battery	6	2	2	226	2.71
SVIB makes no contribution to Reference Battery	8	2	3	233	1.89

*$p < .05$

Table 5. Psychomotor Measures

	Pass/Fail (N = 150)	FTD/Other (N = 150)
Test 1 - X_4	-.19*	.27*
Test 1 - X_5	-.20*	.29*
Test 1 - Y_4	-.14	.27*
Test 1 - Y_5	-.20*	.30*
Test 2 - X_4	-.21*	.27*
Test 2 - X_5	-.20*	.25*
Test 2 - Y_4	-.24*	.26*
Test 2 - Y_5	-.18*	.26*
Test 2 - Z_4	-.15	.25*
Test 2 - Z_5	-.19*	.28*

	Pass/Fail (N = 234)	Pass/FTD (N = 201)
Test 1 - $X_4 + X_5$	-.19*	-.15*
Test 1 - $Y_4 + Y_5$	-.17*	-.16*
Test 2 - $X_4 + X_5$	-.16*	-.16*
Test 2 - $Y_4 + Y_5$	-.10	-.12
Test 2 - $Z_4 + Z_5$	-.12	-.06

*$p < .05$

Table 6. GAT Factors from Study I

Factor	Title
I	Heading
II	Bank
III	Altitude
IV	Traffic Pattern Side Slip
V	Traffic Pattern Bank
VI	Traffic Pattern Position

Table 7. GAT Factors - Study I

Factors	Pass/Fail (N = 140)	Pass/FTD (N = 117)
I	.18*	.03
II	.27*	.37*
III	.00	.08
IV	.15	.09
V	.20*	.13
VI	.04	.08

*$p < .05$

Table 8. GAT Factors - Study II

Factors	Pass/Fail (N = 116)	Pass/FTD (N = 99)
I	.18	.18
II	.15	.16
III	.20*	.25*
IV	.16	.14
V	.06	.16
VI	.20*	.21*

*$p < .05$

Table 9. Simple GAT Scores - Study I

GAT Variable	Pass/Fail (N = 140)	Pass/FTD (N = 117)
Average Pitch Angle Deviation	-.26*	-.18
Average Bank Angle Deviation	-.28*	-.33*
Average Side Slip Deviation	-.19*	-.11
Average Heading Deviation	-.27*	-.19*
Average Altitude Deviation	-.20*	-.22*

*$p < .05$

Table 10. Simple GAT Scores - Study II

GAT Variable	Pass/Fail (N = 116)	Pass/FTD (N = 99)
Average Pitch Angle Deviation	-.28*	-.37*
Average Bank Angle Deviation	-.27*	-.26*
Average Side Slip	-.15	-.19*
Average Heading Deviation	-.09	-.14
Average Altitude Deviation	-.18	-.22*

*$p < .05$

Table 11. Multiple Correlations of Simple GAT Scores

Criterion	N	No. of Predictors	R
Pass/Fail	140	5	.32
Pass/FTD	117	5	.37

Table 12. Cross-Validated Multiple Correlations

Criterion	No. of Predictors	Study I N	Study I R	Study II N	Study II R
Pass/Fail					
Gat Factors	6	120	.42	99	-.08
Pass/FTD					
GAT Factors	6	98	.48	99	.27

Table 13. Multiple Correlations of Simple GAT Scores Corrected for Shrinkage

Criterion	N	No. of Predictors	R	R_c
Pass/Fail	140	5	.32	.28
Pass/FTD	117	5	.37	.33

Table 14. Joint Contributions - Study I

Criterion	Predictors	No. of Predictors	N	R
Pass/Fail	Psychomotor & Gat Factors	11	120	.49
	Psychomotor & SVIB	8	218	.29
	Psychomotor & OBAS	7	210	.30
	Psychomotor & Ref Battery	26	224	.32
Pass/FTD	Psychomotor & Gat Factors	11	98	.51
	Psychomotor & SVIB	8	186	.25
	Psychomotor & OBAS	7	179	.24
	Psychomotor & Ref Battery	26	192	.37

PSYCHOMETRIC SUPPORT FOR ITEM TYPES USED IN A NEW WRITTEN EXAMINATION FOR ENTRY-LEVEL FIREFIGHTERS

Lois C. Northrop

This paper briefly discusses the psychometric background of, and the item types chosen to measure, the six ability constructs represented in a new entry-level examination for use by the D.C. Fire Department. Documentation of this nature is but one phase in the process of developing a new measuring instrument. Extensive task and duty analyses, the linkage of knowledges, skills, abilities and other worker characteristics to these task analyses and the development of the actual test plan are other phases of the entire process which will not be dealt with here.

The six constructs discussed here are the six which were determined to be the most critical and measurable of the 19 cognitive abilities studied.

Since human cognitive abilities are correlated, three to five of these abilities will account for almost all of the variance in performance that can be measured. The addition of more ability measures does not appreciably increase the predictive validity of the test, i.e., the ability of the test to predict future training and job performance. Therefore the six critical abilities selected for inclusion in the entry-level firefighter examination are, for all practical purposes, representative of all the requirements for job success.

The six critical abilities are defined as follows:

1. The ability to read and understand written materials and instructions.

2. The ability to understand and follow spoken instructions or orders.

3. The ability to use simple mathematical formulas or equations.

4. The ability to discover general rules or principles from specific situations or events (learning from experience).

5. The ability to recognize or identify problems or potential problems.

6. The ability to make judgments and decisions when information is incomplete or conflicting.

A search of the psychological literature provided six matching ability constructs which have been consistently identified in numerous factor-analytic studies over the past 30 or 40 years. Many of these studies represent the foundation research in the field of psychometrics. There are differences, of course, among the studies due to the use of a variety of subjects varying in age as well as in ability level. The stability of these ability constructs has persisted despite the diversity of conditions in the factor-analytic studies in which they have been identified.

The first ability - the ability to read and understand written materials and instructions - is represented by the well-known Verbal or Verbal Comprehension factor which, according to the 1976 manual of factor-referenced cognitive tests by Ekstrom, French and Harman of the Educational Testing Service, can be found in at least 125 published studies. The factor is a very stable one. It repeatedly appears in factor analytic studies when a variety of tests in the verbal medium are included in the test battery and it shows great resistance to breaking up into subfactors.

It has been suggested that multiple choice vocabulary tests based on synonyms are the best for picking up individual differences in verbal comprehension. However, assessing verbal ability with isolated vocabulary words alone tends to suggest that Verbal Comprehension is only a subfactor of a broader factor which involves reading comprehension, verbal analogies, matching proverbs, grammar and syntax. Reading Comprehension was chosen to measure verbal ability on the entry-level D. C. firefighter written examination because it represents a more diversified aspect of the ability and has identified a Verbal factor along with a number of other tests of a verbal nature, (analogies, proverbs, grammar). In addition, Reading Comprehension items seemed most appropriate for assessing what was required - the ability to read and understand written materials and instructions.

During World War II, the Army Air Force research psychologists used tests of Reading Comprehension to assess verbal ability in the Air Force Classification Batteries. Items were based on paragraphs which were "simple descriptions of Air Force jobs, the training involved and the individual characteristics required for success." Several questions were asked about each paragraph based either on specific information contained in it or on inferences which could be drawn from the material presented.

Reading comprehension items for the entry-level D.C. firefighter written test consist of very short paragraphs or parts of paragraphs taken directly from the official fire department manuals, training guides and other written materials used in the work of the entry-level D.C. firefighter. These items sample directly the reading material which entrylevel firefighers deal with both during training and subsequently on the job. Such items test their ability to read, understand and interpret this material. Only material of an introductory or very general nature is selected. Some editing may have been done to eliminate any technical wording, which an entry-level applicant would not be expected to be familiar with, or to clarify the statements. No changes are made in the concept or meaning of the information or in the general level of difficulty of the passages to be read and comprehended.

The second ability - that of understanding or following spoken instructions or orders - was identified with an Integration factor in Army Air Force research during World War II. The factor was characterized by the ability to adapt quickly to new instructions and carry them out successfully. Items in the test presented modifications or variations in the instructions given at the beginning of the test.

The same factor was identified as Attention by Wittenborn, the assumption being that attention is required when following a series of rapidly given oral directions, none of which is difficult by itself, but when several are given at once, considerable effort is necessary in order to follow all simultaneously. Tests of attention were designed so that performance on them was not dependent on intellectual ability. Tasks presented were independent of knowledge and content, and involved material familiar to anyone, i.e., digits or letters of the alphabet. Scores depended on the ability to continually sustain mental effort or "concentrate."

The entry-level firefighter test for this ability provides a special action sheet which in addition to letters and digits contains a number of simple geometric shapes. Each item consists of a set of oral directions such as: Make a cross in the first circle and also a figure 1 in the third circle."

When the instructions for an item are complete, the examinees are given from 5 to 10 seconds (depending on the complexity of the instructions) to respond before the instructions for the next item are presented.

Instructions for each item in a test of this nature can be given considerable variety and a wide range of intricacy and yet no single instruction is difficult by itself. The content material (digits, letters of the alphabet, simple geometric shapes) is familiar to all. In this test, an examinee's ability to comprehend and adapt quickly to changing instructions is well tested. This particular test has been in use in the U.S. Civil Service Commission for many years and is considered to be one of the best measures available for measuring the ability to understand and follow spoken instructions or orders.

The third ability - the ability to use simple mathematical formulas or equations - is best represented by the clearly defined Number factor found in over 80 studies. This factor is not a major component in mathematical reasoning or higher mathematical skills. It is simply the ability to perform basic arithmetic operations with speed and accuracy.

The best reference tests for the Number factor are those with the greatest amount of number handling, i.e., tests of the four arithmetic operations (addition, subtraction, multiplication and division). Such tests are outstanding in purity (i.e., they do not load on other factors) and in the size of their loadings on the Number factor.

The item type which will be used in the D.C. entry-level firefighter written test to assess the ability to use simple mathematical formulas is very specifically job related. A comprehensive study of the mathematical abilities required to perform the entry-level firefighter job in D.C. was undertaken and the appropriate level of difficulty for these required abilities was determined. Simple formulas taken directly from the training manuals are presented with numerical values which may be substituted in the formula in order to calculate some quantity. The formulas, of course, involve the four arithmetic operations and facility with the operations determine how easily and quickly an applicant can arrive at the correct answer. Such items, of course, are not as pure a measure of numerical

ability as simple arithmetic problems, since the additional task of substituting a number for the appropriate letter in a formula is required. However, the format of the proposed items is more suitable for assessing the ability actually required in the entry-level firefighter job.

The ability to discover general rules or principles from specific situations is best represented by the Induction factor which appeared as one of Thurstone's 16 primary mental abilities and has continued to appear in the literature. The factor fits the concept of inductive reasoning well, namely reasoning from the specific to the general.

The typical induction test presents groups of words, letters, numbers or figures and the examinee is asked to discover a principle or rule used in their makeup. Using the discovered principle or rule, the examinee can select the appropriate addition or continuation for the group.

The identification of Induction as a factor separate from Deduction and/or General Reasoning is not as straightforward as that for the Verbal and Number factors. There exists some controversy as to whether individual aspects of the reasoning process are separately measurable. Nevertheless a inductive ability throughout the psychometric literature. One of the most enduring of these is Letter Series. Such items consist of a series of letters which follow a pattern. The examinee's task is to discover the rule or principle underlying this pattern and thus find the next letter in the series. Because the Letter Series item type is well documented in the psychometric literature as a measure of Induction it was chosen for inclusion in the D.C. Fire Department entry-level written examination. This item type has been used in other written examinations developed by the Civil Service Commission and has been found very satisfactory for assessing inductive ability.

The ability to recognize or identify problems or potential problems is identified with a factor which has been defined as Sensitivity to Problems. The factor's appearance has been confined to Guilford's laboratory, where it has been isolated in a number of different studies.

The item type chosen for the problem identification items on the entry-level firefighter written examination presents a plan or a situation; the examinee's task is to determine the chief problem with, or the most serious defect in, the given plan or situation. The item type is complex - that is, it has factor loadings on factors other than Sensitivity to Problems.

Although subject matter for faulty plans or futile activities is drawn from common situations which can have been experienced or heard about by anyone, the task of finding the correct answer involves analysis and evaluation of a situation and then thought beyond the given situation in order to generate reasons why the plan or action is faulty and will not lead to the desired result. For these reasons the item type was considered the most appropriate measure of the ability to recognize or identify problems or potential problems.

The last of the six abilities tested in the entry-level firefighter examination - the ability to make judgments and decisions when information is incomplete or conflicting - is best represented by the Judgement factor which was first isolated in several studies, most of them carried out by the Army Air Force Aviation Psychology Research program during World War II. Typically judgment items present a problem requiring a best solution. Not all of the facts bearing on the solution are given and some reasonable assumptions or guesses must be made by the examinee as to what the most likely of several possible occurrences might be. This concept of judgment is obviously complex.

Any suitable item type to assess judgment will have variance in common with other verbal tests but the crucial task in measuring judgment is the supplying of additional data (general knowledge, experience) by the competitor in order to solve the problem.

A review of possible item types suggested the Wechsler (WAIS) Comprehension Item as the best single measure of judgment; such items present "common sense" questions, successful answers to which depend upon one's fund of practical information and one's evaluation of past experience. Questions are of a type which adults have had to answer for themselves or have heard discussed.

Other item types to measure judgment have been of a similar nature and were used in the Air Force research program. Items presented problems of a "common everyday" type, some having to do with work planning, the solution to which rested not upon logical reasoning grounds but on the ability of the examinee to draw upon his "common sense," experience and general information background.

Acts of judgment are necessarily a part of many intellectual functions. Judgment tests will therefore be factorially complex but nevertheless, judgment is an important and assessable aspect of human intellect, necessary to practical success in many occupations including that of firefighting.

Judgment items for the entry-level D.C. Firefighter written examination draw upon knowledge or experience of a very general, non-technical, and current nature, the level of which is based on that for any high school graduate. A five-alternative multible choice format is used. A short stem is sought in order to reduce the reading comprehension variance of the question. Distractors may be based on false, irrelevant or incomplete assumptions; the correct alternative should have a documented source and should not be predicated on mere opinion. Such items will appropriately measure the required ability.

A METHODOLOGY FOR ASSESSING PREFERENCES IN SPATIAL ARRANGEMENTS

Jim Ingram Martin
Navy Personnel Research and Development Center
San Diego, California 92152

Gail Metzger and Colleen L. Zelazny
San Diego State University

INTRODUCTION

The design of living spaces has a strong impact on the "quality of life" perceived by those who live and work within those spaces. For the Navy and other organizations which control the physical surroundings of their personnel for prolonged periods of time, there is an increased awareness of the importance of structuring the environment to meet the needs and desires of its occupants. The habitability program of the Navy has looked at a wide variety of means of improving the perceived quality of living spaces. Such a program is recognized as being desirable for the sustained physical and psychological well-being of those who are in the Navy. With the advent of the all-volunteer force, improved life spaces are recognized as being essential in order to both attract and retain the personnel the Navy needs.

Many factors have been examined by habitability studies. Some of these factors have involved light levels, color schemes, surface treatments, noise levels, temperature variations, and ventilation rates. An area which has not received much attention has been that of spatial arrangements. The studies that have been conducted in this area have primarily involved the redesign of work spaces. In those analyses, the arrangement of physical space to meet time and energy requirements, not personal desires, has been the primary objective. For the life spaces, no effort has been made to examine optimal spatial arrangements.

A prime requirement for developing improved habitability programs is the determination of spatial preferences. A second reason for being concerned with the assessment of spatial preferences is the possibility of differences in the ways in which men and women perceive and utilize space. Many architects, both male and female, profess that such sexually-distinguishable differences exist. For the Navy, with the prospect of sexually-integrated crews, differences in the spatial preferences of men and women would impact the effectiveness of habitability improvement programs.

A study of the attitudes of men and women crew members aboard the USS SANCTUARY led to the observation that "women aboard ship express a greater need for personal privacy than do the men" (Martin et al., 1973, p. 62). Whether this expression was simply a verbal phenomenon, or whether it reflected some stronger, behaviorally-based need, could not be ascertained at the time of that earlier study.

The primary objective of the research reported in this paper was to develop tools and techniques which could be used for determining preferences

in spatial arrangements. A secondary objective was to administer these tests to a limited sample of men and women to determine whether conspicuous differences in spatial preferences existed between the two sexes.

METHOD AND RESULTS

After examining a number of means of determining spatial preferences, three sets of assessment tools were developed. These were a layout diagram, a figure selection test, and a questionnaire. For this exploratory effort, the life space selected as the basis for the study was a dormitory room shared with one other person of the same sex.

The first tool, the two-dimensional reduced-scale layout diagram, permitted respondents to move cutouts depicting two twin beds and two wardrobe/dresser units in a 9 x 12 foot room. The second assessment tool presented sets of four possible room arrangements (derived for the same furniture and space as for the first test) which the subjects were to rank in order of preferences. The third tool was a questionnaire obtaining responses to verbal statements about preferences in layout design as well as to the need for privacy.

LAYOUT DIAGRAM

A preliminary experiment was conducted with the layout diagram in which it was given to 53 male subjects ranging in age from the late teens to early twenties. With only four items of furniture and with a relatively small space, it would seem that the varieties of responses would be limited. In fact, however, a considerable number of different responses was possible. Considering only the positions of the two beds against the walls, there were ten basic possible configurations. If one or both of the beds were moved out from the walls, more variations were possible (but these were counted as variants of the basic layouts). The introduction of the two wardrobe units greatly increased the number of designs. A total of 33 basic variations for beds and wardrobes was found to exist.

Considerable individual differences in preferences for spatial arrangements were found in this preliminary experiment. Of the ten possible basic configurations based on bed positions alone, eight were selected by the respondents. While 40 respondents chose basic configurations with both beds touching the walls at corners of the room, 13 respondents chose to have at least one of the beds free-standing. Of the 33 possible configurations based on bed and wardrobe positions, 18 were selected by the respondents.

Despite the diversity of the responses, some general findings could be promulgated. Fully 70 percent of the respondents arranged their beds so that they were parallel with their roommate's. More than 77 percent positioned their beds so that they were in corners of the room. Slightly over 62 percent could be accommodated by a selection from three basic designs.

Following this pilot study, the layout diagram test was modified and presented, in conjunction with the figure selection and questionnaire tests, to a sample of ten males and ten females. This limited sample group also differed from the first in that the respondents were primarily in their 20's and 30's.

Considerable diversity in responses was found. Of the ten configurations based on bed position alone, six were selected. Of the 33 configurations based on bed and wardrobe positions, 11 were selected.

Nevertheless, some uniformity of response was observed. One particular bed position configuration was selected by 40 percent of the respondents. In fact, one specific bed and wardrobe configuration was selected by 30 percent. Beds were positioned against the walls, in corners of the room, by 90 percent. The preference for locating the beds parallel to one another, observed in the pilot study, was not upheld in this one; the selection of parallel versus perpendicular orientations was made by exactly 50 percent of the respondents.

These differences in responses between the pilot study and this later experiment may have been due to the increased age, and, hence, maturity and accumulated experience of the respondents in living in dormitory conditions or in visualizing space from two-dimensional cutouts.

The most surprising finding was the reduction in diversity of responses and the concentration of such a large percentage of respondents on a single configuration.

SCORING OF LAYOUT DATA. The tabulation of results from the layout data proved unexpectedly difficult. The preliminary studies had indicated some of the types of data which could be collected. The room designs generated certain impressions which were difficult to define. Efforts to stipulate definitions almost inevitably ran afoul of exceptions. Nor were reliable instruments available to make all the measurements. A planimeter for measuring surface area gave discrepant results each time it was used. Nevertheless, some 16 different items were developed which could be consistently measured. These are listed in Table 1 and a diagram of the room and its furnishings is depicted in Figure 1. The items on which data were taken included factors relating to the orientation of the beds, the distances from the beds to other objects, the visibility of other objects from the bed, and the personal area defined by each bed and wardrobe combination.

The analysis of the data obtained from the men and women indicated remarkable overall similarities. Table 1 presents the averages of the responses obtained from each sample. The greatest difference between the men and women appeared to lie in the division of the room into personal areas. Whereas the men generally divided the area equally, the women tended to divide the area unequally (70 percent versus 30 percent). Nevertheless, the unequal distribution of area by both the men and the

Table 1. Layout Diagram Results

CATEGORY AND ITEM	MALES	FEMALES
Bed Orientation:		
1. Parallel	40%	60%*
2. Perpendicular	60%	40%
3. Converging view	100%	100%
4. Diverging view	0%	0%
Bed Distances:		
5. Pillow separation	10.8'	10.0'**
6. Bed separation	4.0'	3.9'
7. Respondent to door	8.9'	9.9'
8. Roommate to door	8.7'	8.0'
9. Respondent farther from door	50%	80%
Bed Visibility:		
10. Respondent to roommate	80%	80%
11. Respondent to door	80%	70%
12. Roommate to door	70%	80%
Personal Area:		
13. Respondent's frontage (bed to ward.)	6.[illegible]'	5.4'
14. Roommate's frontage (bed to ward.)	5.6'	5.8'
15. Equal frontages for both roommates	60%	30%
16. Equal areas for both roommates	70%	30%

* Percentages indicate respondents arranging their room layout to include the listed feature.

** Measurements indicate the average distance in feet.

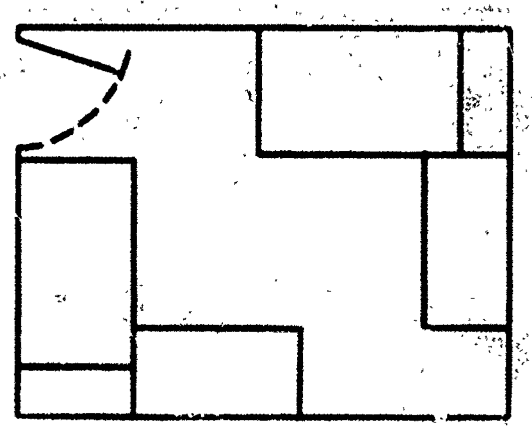


Figure 1. Room Layout Diagram

women was not done for personal gain; in both cases, the roommate benefited more frequently than the respondent.

FIGURE SELECTION TEST

The figure selection test was designed to ensure that the respondents would consider the major alternatives possible with the room layout design. It also attempted to separate out the component decision points involved in arriving at a satisfactory design.

The test consisted of sets of four possible configurations. Within each set, the respondent was required to rank order his preferences. The sets were designed so that, inasmuch as was possible, one primary factor was being varied while the others were held constant.

The first two sets consisted of the room with two beds in different arrangements. The locations of the pillows on the beds were not indicated and the wardrobes were not included in these sets.

The first set offered the choice between beds which were parallel and beds which were perpendicular. Both the males and females were equally divided between parallel and perpendicular arrangements.

The second set offered the choice between beds which were positioned in different corners of the room at increasing distances of separation. Both the males and females prefered the greatest separation (with beds in opposite corners) and avoided the least separation (parallel beds occupying adjacent corners on the shorter wall).

The third and fourth sets consisted of drawings of the room with two beds on which the locations of the pillows were indicated (the wardrobes were not included in these sets).

The third set offered the choice, with beds arranged in parallel at opposite corners of the room, of having the heads of the occupants facing in converging, parallel, or diverging directions. Near universal agreement was obtained favoring the converging arrangement. This preference was followed, in descending order, by the parallel arrangement with both roommates looking into the hall; by the same arrangement with both looking at the outside wall; and by the diverging arrangement.

The fourth set offered the choice, with the beds in perpendicular arrangements at adjacent and opposite corners, of having the pillow positions at varying distances of separation. Again, near universal agreement was obtained favoring the maximum pillow separation, with the heads of the beds in opposite corners of the room. The rank ordering of this and of the remaining preferences indicated that two factors were at work: One factor involved keeping the heads of the beds against walls (i.e., in corners of the room); the other involved attaining maximum separation within that constraint.

The fifth and sixth sets of drawings complicated the selection process still further by the addition of the wardrobe units.

The fifth set offered the choice, with the beds in perpendicular arrangement at opposite corners, of locating the wardrobes at various positions with regard to the heads of the beds. Both males and females prefered to have the wardrobe positioned so that there was a maximum amount of open area. They least prefered selections which closed off an area of the room from visibility from other areas.

The sixth set identified one of the beds as that of the respondent and offered the choice of having it be visible or hidden from view from the hall. The responses of the males and females showed little consistency with this set. In general, a slight preference appeared to exist for the more open floor plans, but the results were so highly individualized that no general conclusions can be supported with any degree of confidence.

The figure selection test revealed several findings that were also reflected in the room layout diagram. Respondents had little preference between parallel and perpendicular bed arrangements. They prefer to maximize the separation between beds. The heads of the beds must be positioned against a wall and respondents will accept a lesser separation between beds in order to have the heads in this position. The positions of the heads should be such that roommates are facing in a converging direction; parallel viewing arrangements are not prefered, but, if they must be selected, the respondents would require that they be oriented to afford a view of the entry to the room. A preference for maximizing the amount of open area in the room appeared present in most responses. Designs with sections of the room visually closed-off from the other were not regarded favorably. When respondents were given the opportunity to shield their beds from view from the hall, they showed no preference for selecting those arrangements.

QUESTIONNAIRE

The questionnaire was designed to tap the thought processes which might be underlying the room layout and figure selection test behavior. Because it was a written test which required the verbalization of a response (as opposed to positioning the cutouts or selecting a design), it was anticipated that the respondents might exhibit more social constraints in developing their answers.

The questionnaire asked specific questions about the design process and some general ones about the need for privacy.

DESIGN PROCESS. One portion of the questionnaire consisted of a series of items related to the various identifiable aspects of the physical/psychological arrangement of the rooms. These questions were related to the discernable features which were being measured on the layout diagram.

The responses to the questionnaire were compared with the room layouts designed by each respondent. This was done in order to determine whether

the verbalized expression was realized in a physical manifestation; in other words, to find out whether the questionnaire and the layout would indicate the same response. (This had been an issue because annecdotal evidence had suggested that the verbal behavior may be quite different from the actual physical behavior.)

In general, the responses of the subjects tracked their actual layout diagrams fairly closely. In some cases, tradeoffs were necessitated by the room design and not all idealized responses could be accomodated. In most instances, however, the verbal expression about design characteristics was paralleled in the actual behavior.

The one major exception occured with the responses to the question of whether the respondent would consider it desirable for both roommates to have equal amounts of personal area. All respondents, male and female, replied affirmatively, but, as was noted earlier, 30 percent of the men and 70 percent of the women did not achieve that goal in their layout diagrams.

PRIVACY. One portion of the questionnaire was directed at determining the sensitivity of the respondents to the need for privacy. This was assessed in two ways. A direct assessment was obtained by asking the respondent to indicate, for each of nine activities performed in the room, his feelings of a need for privacy on a 5-point scale. An indirect assessment was obtained by asking, for the same activities, whether or not the door would be open or closed. Responses to these two questions (presented in Table 2) were used to determine an overall level of sensitivity to the need for privacy for each respondent.

The responses of the subjects to the two types of questions tracked one another closely, supporting the belief that both questions were directed at the same basic area. There was a confounding of results by some additional factors, however. For example, with regard to listening to a radio, a number of subjects indicated that they would close the door even though they expressed no particular need for privacy (they would close the door out of consideration for their neighbors).

In general, the females indicated a greater feeling of a need for privacy than the males. (The one exception to this generalization was with regard to writing where a greater number of males than females would close the door.) These findings were significant at the 0.025 level.

The privacy rating was used to divide the sample of men and of women into additional subgroups of high and low scores. The data obtained with the layout diagram, the figure selection test, and the design process portion of the questionnaire were then examined to determine whether there were any significant differences in the responses of those experiencing different levels of need for privacy. Despite the numerous measurements and comparisons, no conspicuous and consistent results were apparent.

Table 2. Privacy Questionnaire Results

ACTIVITY	PRIVACY* Male	PRIVACY* Female	PRIVACY* Difference	DOOR POSITION** Male	DOOR POSITION** Female	DOOR POSITION** Difference
Sleeping	4.3	4.7	0.4	90	100	10
Napping	3.6	4.5	0.9	80	100	20
Reading	3.3	3.6	0.3	80	80	0
Dressing	3.3	4.5	1.2	80	90	10
Grooming	2.4	3.9	1.5	60	90	30
Writing	2.8	3.3	0.5	80	70	-10
Talking	2.3	3.3	1.0	40	60	20
Listening	1.6	2.2	0.6	50	50	0
Relaxing	2.0	2.5	0.5	20	40	20

* Rated on a scale from 5 to 1 of decreasing need for privacy.
** Scored as percentage of respondents who would close the door while engaged in the listed activity.

CONCLUSIONS

DESIGN CHARACTERISTICS

The results of this study indicate that, although there are a great many individual differences in terms of preferences for the way in which a room will be arranged, there are some basic design characteristics which will be more appealing to the greater number of people. Such factors as having the beds facing in convergent directions, positioning beds so that the heads are against the wall, maximizing the separation between the heads of the beds, etc., are prime considerations. Other factors such as orienting the beds in perpendicular or parallel fashion are not especially important.

A major finding of this study is that, although there are significant differences in the expression by males and by females of the need for privacy, these differences are not reflected in sexually-differentiated preferences for spatial arrangements. Both the males and females demonstrated substantially similar responses to the layout diagram and to the figure selection test.

If substantiated by further research with additional subjects, and if upheld for other life spaces which would need to be examined, these findings would allay the concerns of habitability engineers about designing life spaces which may prove equally satisfactory to male and to female personnel.

ASSESSMENT TOOLS

This study has also indicated that a battery of assessment tools may be used to determine spatial preferences for life spaces. The results obtained with such varied tools as the free-response, layout diagram; the forced-choice, figure selection test; and the verbally-oriented questionnaire, all led to similar conclusions.

The layout diagram permits maximum variation in response, yet is limited because the subject may be constrained by his ability to conceptualize alternative patterns. The figure selection test permits particular design factors to be isolated by holding certain factors constant while manipulating others. The questionnaire permits an assessment to be made of internal factors, but care must be taken to distinguish between those questions pertaining to tangible, physical aspects of design and those pertaining to the emotional, psychological aspects or feelings.

Each test has its own particular merits. A battery of tests such as was developed for this study could be used to determine the spatial preferences of subjects for the improved habitability design of their life spaces.

REFERENCE

Martin, J. I., Sabeh, R., Kritz, A. C., and Driver, L. I. Women aboard ship: Habitability design considerations. Naval Electronics Laboratory Center Technical Note, TN 2418, 26 June 1973.

OBVERSE FACTOR ANALYSIS TO ENHANCE THE VALIDITY OF SELECTION TESTS*

Benjamin B. Weybrew
Head, Personnel Assessment Division
Naval Submarine Medical Research Laboratory
Naval Submarine Base New London
Groton, Connecticut 06340

Fourteen years ago I was privileged to be the keynote speaker at the 6th annual MTA conference held at the Coast Guard Institute in Groton, Connecticut. Entitled "Testing is Serious Business", my address contained several examples of my discontent with certain test methodology in vogue at the time. One of these problem areas I mentioned was the relative usefulness of configural versus linear or summative models in scoring our tests and in deriving our predictive formulae. This paper then becomes a modest attempt to substitute action for talk by providing a demonstration of what happens to the validity of selection tests whose scores were derived from weighted patterns of responses rather than weighted or unweighted sums of responses, the latter being the commonly accepted scoring technique used by most test practioners.

The need for a new approach to submariner selection was based upon several observations regarding certain aspects of the submariner selection data collected over the past two decades. In the first place, involving a variety of aptitude, interest and personality tests, our linearly-derived, validity coefficients, single or multiple, rarely exceed 0.40 with our available training criteria. Secondly, in an era of shortages of quality submariner volunteers when false positive selection errors (reject a good candidate) are very serious, the observation that there are almost an infinite number of trait configurations descriptive of the 30-40% of the submariner candidates who fail at various phases of their career. Aptitude deficiencies, deficient or inappropriate motivation, attitudes or interests, and emotional instability in an endless variety of patterns characterize most submariner "attrites" (Weybrew, 1963). It seems much too simple, certainly unrealistic, to expect a linear, summative equation made up of one or more aptitude scores for example, to yield substantive predictive indices with respect to these attrition criteria.

Thus, stated more precisely, the objectives of this study were twofold: (1) Empirically, to identify personality types characterizing "attrites" and successes, then (2) to investigate the differences

*The opinions or assertions where they appear in this paper are those of the author and are not to be construed as the official views of the U.S. Navy Medical Department.

in the predictive validity of standard selection tests for "good" and "poor" fits to these "types". There was one assumption underlying this study, viz., that variables interrelate differently within different classes or types of individuals. It follows, therefore, that only when the test battery contains a single common factor and the criterion population is homogeneous in the sense that the interaction matrix of the variables (items of a questionnaire, in this particular study) are equivalent for all criterion sub-groups, will a given prediction formula be maximally effective for the total criterion population. It was therefore the burden of this study to show that obverse factor analysis is one possible technique for grouping the total criterion population into homogeneous sub-groups within which differential predictive validity may be found.

METHOD AND PROCEDURE

As alluded to in the introductory comments the method of choice for deriving personality "types" was obverse or inverse factor analysis (Cattell, 1952), a factorization technique by which hypothetical factors are derived from between-person covariance or correlation matrices. Found early in the history of submarine psychology to be useful in identifying trait patterns characterizing successful submariners (Weybrew, 1953), this factor analytical technique (called Q-technique if Q-sorts are involved) yields factors whose structure is defined by factor loading on persons rather than loadings on tests or measures produced by the more common between-test or R-technique of factor analysis.

Subjects. The sample upon which the typology derived by obverse factor analysis was based consisted of twenty enlisted candidates for submarine training. This sample was drawn randomly from a population of 800, 15% (3 Ss) from high-achieving successes (stanine 9), 15% from the stanine "1" success group and the remainder (14 Ss) from the attrition segment. The validation samples consisted of two independent Submarine School classes, N=277.

Data collection and analytical techniques. A psychiatric screening questionnaire, "custom-tailored" for submariner candidates, the Personal Inventory Barometer (PIB), was used to identify the adjustment types (Weybrew and Youniss, 1957). Validated mainly to identify adjustment failures at the training level, the 52 keyed items of the PIB yielded high concurrent validity with such well-known personality tests as Scale 7, Psychasthenia, MMPI (0.73, N=250) and with the Guilford-Martin (GAMIN) -0.66 and -0.73 with the Absence of Inferiority and the Nervousness factors respectively. The PIB employs a 10-point multi-category response format extending from "0", "Not at all like me" to "9", "Exactly like me" with 3 intervening anchor-categories.

The analytical procedure consisted of the following steps: (1) The 52-item array for each of the 20 Ss was first converted to ipsative form, that is deviation scores (stanines), were calculated from the means and SD's for each of the 52-item response arrays for each of the 20 Ss; (2) Thurstone's Group Centroid Method of factor analysis was applied to the 20x20 inter-person matrix of Pearson Product Moment coefficients and (3) the resulting 6x20 factor matrix rotated orthogonally by a simple geometric technique (Fruchter, 1954). The rotational criterion was simple structure. The "person-vectors" were coded but not identified during the rotational procedure so as to provide a "blind" control on the person (the author) carrying out the calculations.

RESULTS

Delineation of the Obverse Factor Types. Thirty-six per cent of the between-person correlation coefficients were significant at the 5% confidence level. Nevertheless, the correlations were low, with a mean coefficient of 0.23 and a S.D. of 0.12. For the most part, the residual matrix was exhausted after the extraction of six reference centroids although Tucker's criterion was met after the fifth residuals were computed. The mean and S.D. of the residuals were .07 and .05 respectively. These six reference axes were then rotated orthogonally until simple structure was approximated. The factor matrix is presented in Table 1. (Table 1 on following page)

TABLE 1

Orthogonally Rotated Factor Loadings of

Twenty Enlisted Men*

Subject Code Number	Obverse Factors						
	I	II	III	IV	V	VI	h^2
1	.40	.35	.36	.38	-.09	-.02	.56
2	.23	**.58**	.06	-.04	.09	-.13	.42
3	.31	**.59**	-.24	.15	.23	.06	.58
4	-.02	.06	.34	.11	**.72**	.00	.65
5	.37	.20	-.07	.28	.08	.00	.27
6	.42	-.24	**.54**	-.19	.43	-.15	.77
7	**.65**	-.27	-.04	.34	.18	.00	.64
8	.29	-.25	.10	.16	.11	.31	.29
9	.47	-.26	.15	.08	.25	.03	.38
10	**.56**	.23	-.05	.18	-.28	-.05	.48
11	**.66**	.14	.17	**.46**	-.14	-.09	.72
12	.12	.01	.16	.06	**.84**	.00	.75
13	.08	.00	.12	**.82**	.02	.00	.69
14	.13	-.06	.19	.13	**.70**	.17	.59
15	.23	.12	.16	-.14	.30	.00	.20
16	**.62**	-.10	-.02	.15	.28	-.21	.54
17	**.52**	.30	.15	-.22	.23	.00	.48
18	**.55**	-.14	.02	-.01	.09	.10	.34
19	.17	-.12	**.87**	-.09	.36	-.02	.94
20	.35	.35	.29	.21	.37	-.02	.51

*Factor loadings which are underlined indicate the persons used to identify the factor.

An inspection of the underlined loadings in Table 1 indicates that all of the persons except those with code numbers 1, 5, 8, 9, 15 and 20 were used to identify some one or other of the factors. Judging from the low communalities (h^2 in Table 1), it is apparent that persons with code numbers of 5, 8, and 15 had item profiles unique to this group. The remaining three out of the six persons were rejected as identifying persons largely on the basis that the vectors representing their item profiles failed to rotate into any of the factor hyperplanes.

The persons identifying each factor in the rotated solution in Table 1 appeared to cluster reasonably well with respect to the criterion groupings used to select the 20 subjects for the obverse factor analysis. Thus, Factor I had the highest loadings by four "academic" failures and two "temperamental" failures. Factor II, on the other hand is a doublet, loaded by two men who graduated in the upper stanine. Factor III is also a doublet loaded by one low achiever (Stanine 1) graduate and one "academic" failure. Factor IV contains two high-loading "academic" failures; however, one of them (code number 11) also loads Factor I, indicating an overlap between the two factors. Factor V appears to be a triplet loaded by two "academic" failures and one "stanine 1" graduate. Finally, Factor VI appears to be a residual factor since there are no high-loading persons identifiable in the present solution. Thus, in sum, all but one of the five obverse factors were identified by submariners who failed or were low-achievers in basic training. Factor II, on the other hand, is identified by high-achieving Submarine School graduates.

While not directly pertinent to the methodological emphasis of this paper, it should be mentioned for the more clinically-oriented psychologists, that once the obverse factors have been identified by the people with high loadings on each factor, the structure or content of these factors can be examined by a relatively simple procedure. That is, to content analyze the most (and least) descriptive items (in this study the PIB items) making up the protocols for the persons with high loadings on each obverse factor. For example, the PIB patterns suggest the following factor content: Factor I (F_1) nervousness, frustrated, F_2 - good impulse control, socially sensitive, F_3 - cyclothymic, and F_4 and F_5, characterized by rather similar neurotic patterns.

Relationship of fit to "obverse types" and test validity.

The rationale for this, the central part of the study, is that the degree to which the persons within a subgroup fit a "failure" pattern, to that degree will "success" predictors lose validity. Conversely, to the degree the persons within a sub-group fit a "success" type, to that degree will predictive validity be enhanced.

To test this proposition the following procedure was implemented: The PIB was administered to one enlisted Submarine School class of 322 men (none of the 20 men used in the obverse factor analysis were included in this sample). With the item responses in Stanine form, three scoring methods were applied to the PIB items most and least descriptive of the persons defining the obverse factor types. These were: (1) The sum of those items least characteristic of the factor subtracted from the sum of those items most characteristic; (2) the simple sum of those item responses most characteristic; and (3) the ratio of the sum of the most characteristic item-responses to the sum of the least characteristic responses. Thus, three scores were obtained for each of the five factors identified in Table 1.

Since all of the obverse factors except Factor II were associated with submariner "attrites" and Factor II with submariner "successes", the question was raised as to whether any of these scores based upon the personality test configurations delineated by obverse factor analysis could differentiate between the success and failure sub-groups of an incoming class of submariners. Table 2 contains the results of this analysis.

It is an interesting fact that all three keys for F_1 in Table 2 were significantly discriminatory between those who succeed and fail in Submarine School. Moreover, two of the 3 keys were discriminating for F_5. One reason for the validity of these two factor scores quite probably was the greater number of identifying PIB items, 10 for F_1 and 14 for F_5, as compared with 6 items each for the remaining factors. The greater number of items in the factor clusters for F_1 and F_5 quite probably resulted in enhanced reliability of the derived scores.

The final part of this study (and perhaps the most interesting) consisted of an examination of the effects of grouping "good" and "poor" population fits to the criterion types upon the predictive validity of the Navy Arithmetic-Mechanical Aptitude Test with respect to the Submarine School pass-fail criterion. The methodological hypothesis entertained here was that to the degree to which the persons within a population sub-group fit a failure pattern, to that degree will "success" predictors lose validity. Conversely, to the degree that persons within a group fit a "success" type, to that degree will predictive validity be enhanced. Keeping in mind that within the total group the tetrachoric correlation between the aptitude test and the criterion is .40 the data in Table 3 would seem to be suggestive.

Looking at Table 3, it is well to recall that Factors 1, 3, 4 and 5 were all defined by failure types, and F_2 by success types. Looking at F_1 for all scoring keys, it is noted that within the population sub-group, fitting the failure type depicted by the items identifying F_1, the correlation with the criterion, for all keys, is significantly lower than the same predictor-criterion relationship within those persons who fit this failure type. One the other hand, the data for F_2, a success

TABLE 2

Comparison of Factor Scores for 281 Graduates and 41 Failures in Submarine School

Factors	Scoring Key I		Scoring Key II		Scoring Key III	
	t-ratio	p[a]	t-ratio	p	t-ratio	p
F_1	2.72	.005	2.80	.005	3.72	<.001
F_2	0.45	n.s.	0.03	n.s.	0.09	n.s.
F_3	0.21	n.s.	1.1	n.s.	0.22	n.s.
F_4	0.07	n.s.	0.02	n.s.	0.12	n.s.
F_5	1.77	.04	2.40	.01	0.83	n.s.

[a]Probability based upon a "one-tailed" hypothesis consistent with the content of each factor.

TABLE 3

Tetrachoric Correlation Coefficients of Arithmetic plus Mechanical Scores with Submarine School Stanines for "Good" and "Poor" Fits to the Criterion Types Identified by Obverse Factor Analysis (Total N = 277)

		Scoring Key I	Scoring Key II	Scoring Key III
		r_{tetra}	r_{tetra}	r_{tetra}
F_1	High[a]	.38[b]	.39[b]	.39[b]
	Low	.54	.54	.54
F_2	High	.52	.54	.61[b]
	Low	.46	.48	.46
F_3	High	.46[b]	.34[b]	.44
	Low	.61	.57	.50
F_4	High	.57[b]	.38[b]	.58[b]
	Low	.41	.59	.43
F_5	High	.41[b]	.40[b]	.54
	Low	.59	.57	.48

[a]High-Low signifies above and below the approximate median of the distributions of scores derived for each group of items associated with each factor.
[b]Differences between coefficients significant at less than the 5% level (one-sided test).

type, indicate that those persons who fit the success type (scoring key III only) show a higher predictive relationship with the criterion than those who do not fit the success type, or, more exactly, fit the success pattern less well. F_3, also a failure type, shows for 2 of the three scoring keys the same sort of discrepancies in predictive validity that were observed in F_1, that is, if we hold the sub-groups constant, we obtain an increase in validity for those Ss who did not fit the failure classification as defined by the analysis. For F_4, only scoring key II yields a significant difference between the coefficients in the predicted direction. For reasons unknown, scoring keys I and III for F_4 yield differences in the reverse direction. On the other hand, both of the statistically significant keys for F_5 are in the predicted direction.

SUMMARY AND CONCLUSIONS

There were two general questions to be answered by this study: First, is it possible by means of obverse factor analysis to classify persons meaningfully in terms of differences in personality test item configurations? Secondly, having isolated these person-factors, what happens to predictive validity within population sub-groups showing good and poor "fits" to these classes?

There was one fundamental assumption underlying this study, viz., that variables interrelate differently within different classes or types of individuals. It follows, therefore, that only when the test battery contains a single common factor and the criterion population is homogeneous in the sense that the interaction matrix of the variables (items of a questionnaire, in this particular study) are equivalent for all criterion sub-groups, will a given prediction formula be maximally effective for the total criterion population. It was therefore the burden of this study to show that obverse factor analysis is one possible way to group the total criterion population into homogeneous sub-groups within which differential predictive validity may be found. What did the results show?

The major finding is contained in Table 3. The data in this table demonstrated differential predictive validity of the Arithmetic-Mechanical Aptitude scores within "good" and "poor" fits to the adjustment classes identified by the obverse factor analysis. It was shown that within a group of submariner volunteers who fit the "failure" class, the aptitude predictive validity is significantly lower than for those who fit the failure class less well. At the same time the validity of the same predictor within the group fitting the success "type" was higher than was found for the group failing to fit the success type.

These findings would seem to be in accord with expectations. Factorial composition (i.e., what dimensions the test items are tapping) would seem to depend upon the characteristics of both tests and persons tested and

that differential predictability will be found for the sub-groups compounded within the population sample. Actually a different prediction formula is needed for each of the population sub-classes.

In addition to identifying trait clusters descriptive of persons who failed in Submarine School, the results of this study suggested a somewhat different approach to submariner selection (Weybrew and Kinsey, 1968). After applying the selection battery to the volunteer population, the total group should be subdivided into classes according to the goodness of fit to the success or failure trait patterns as defined empirically as outlined in this study. Then, according to the findings of this study, maximal prediction efficiency will be obtained by deriving a separate prediction formula for each population sub-grouping.

In short, it appears that workers in personnel selection (including submariner selection) most probably are consistently underestimating the validity of their predictor measures as a result of a number of heterogeneous groups being included in the population sample. Once the sample has been classified according to factorial composition (and an obverse factor analytic technique is one possible way to do this), predictability within population sub-groupings should be greatly enhanced. Following somewhat the paradigm outlined in this study, more profitably, perhaps, by isolating population sub-classes by means of a more varied configuration of measures, not only psychiatric screening items as used in this study, but also aptitude, interest, and values scores as well as selected items of biographical information, it should be possible to raise single and multiple validity coefficients to the 0.80 - 0.90 range for certain predictive problems.

REFERENCES

Cattell, R. B. 1952. Factor analysis: An introduction and manual for the psychologist and social scientist. Harper and Brothers, New York.

Fruchter, B. 1954. Introduction to factor analysis. New York: D. Van Nostrand Co., Inc.

Weybrew, B. B. 1953. Q-methodology in criterion research. NavSubMedRschLab Rpt. No. 239.

____________, and Youniss. 1957. The personal inventory barometer (PIP) I. Development of the questionnaire. NavSubMedRschLab Rpt. No. 290.

____________ 1963. Psychological problems of prolonged marine submergence. In Burns, N.M., Chambers, R.M., and Hendler, E. Unusual Environments and Human Behavior. Free Press of Glencoe.

____________, and Kinsey, J. L. 1968. A systems analytical approach to assessment of human reliability. NavSubMedRschLab Memo Rpt. No. 68-15.

Improving Estimates of the Standard Error of the Mean

by: Donald O. Shepherd, Robert L. Malone and Jack A. Kavanagh

Loyola University of Chicago

Statistical estimates of the standard error of the mean are apt to be somewhat erroneous when they fail to consider the measurement error, or conversely the reliability, associated with the items, scales, or operations by which such measurement is attempted.

In the case of test items, as this paper illustrates, estimations of the standard error of the mean are more accurate when based on a formula containing a correction for error of measurement.

While estimates of the standard error of the mean have traditionally recognized variance as a function of the sample of people, Peters and Van Voorhis[1], as well as Shepherd and Winiewicz[2] have cited the importance of acknowledging variance associated with the sampling of the test items. As the present paper illustrates, each type of variance, i.e. people or tests, may be isolated under special conditions. For example, Peters and Van Voorhis (p. 134 Formula 66) give as the standard error of the mean in the case of the correlated test samples matched on an infallible criterion the following formula:

1)

$$\sigma_{\bar{x}} = \frac{\tilde{\sigma}_x}{\sqrt{n}} \sqrt{1 - r} \quad ,$$

where r is the reliability coefficient of the test.

Peters and Van Voorhis (pp. 134-135) go on to state: "We would have such matching on a true criterion where the same group was to be retested, for here the paired individuals are the same persons: consequently truly paired as to ability. The variability of the means to be expected if we should repeatedly retest the same group is probably what we usually have in mind when we think of the standard error of a mean; hence formula (1) is the one most often to be used."

Clearly, what Peters and Van Voorhis are using in the above formula for r is the reliability coefficient and that matching on an infallible or true criterion is testing the same group twice, i.e. the test-retest reliability coefficient. The "variability of the means" is a consequence of measurement error as the variability cannot be due to the sampling of people.

[1]Peters, C.C., and Van Voorhis, W.R., Statistical Procedures and Their Mathematical Bases (New York: McGraw Hill, 1940).

[2]Donald O. Shepherd and Casimer S. Winiewicz "Compleat Formula for the Standard Error of the Mean", Proceedings, 79th Annual Convention, American Psychological Association, 1971, pp. 97-98.

Peters and Van Voorhis used the term "standard error of the mean" while using either the formula for the sampling of people (classified formula p. 131);

2) $$\sigma_{\bar{x}} = \frac{\tilde{\sigma}_x}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$$

or for the sampling of test items or forms (measurement error formula p. 134):

3) $$\sigma_{\bar{x}} = \frac{\tilde{\sigma}_x}{\sqrt{n}} \sqrt{1 - r}$$

Formula (2) is to be used for determining the variance of the sample means for a single test form where the thing that varies is the sample of people. Formula (3) is to be used where the sample of people remains constant (i.e. a population), but the sample of test forms (items) varies.

Actually, however, most situations entail a sampling not only of people from some population, but a sampling of items from some domain of test items. Hence, realistically, it is most important to deal with both types of variance simultaneously as provided in the hypothetical example explored in the remainder of this paper.

For example, let us consider a minature representation as follows in which tests as well as people differ from sample to sample:

> The total population (N) is 4 and sample size (n) is 3. The people are represented by the letters A through D. The domain of the test items (K) is 5, sample size (n) is 3, and length of the test form (k) is 3. The sizes of the universes (people and test) are here kept finite for practical and illustrative purposes.

Table I (following) contains the population of people and test items. It also has each individual's response to each item in the domain of test items.

Table II contains every possible sample of people of size 3 and every possible sample of test items of size 3 (test form).

Table III contains the score for each person by test form. At the far right under T is each persons "true" score calculated from the sample of tests. The population mean for each unique test form is the bottom row of scores (μ_f), i.e. population means for each test form calculated across all persons. Finally, across all persons and all test forms the overall population "true mean" (μ_t) is 1.50. Relevant error variances and other data have been computed and are recorded at the bottom of Table III.

Table IV contains the sample means of every possible combination of people (in groups of 3) and test form. For example, the sample mean of 1.67 at the

intersection of Form I and sample ABC was determined by taking the scores that individuals A, B and C earned on those items (1, 2, and 3) comprising test form I. At the far right of Table IV are the true means ($\bar{T}$) of each of the samples. Each true mean is the avearge of the sample means across all possible test forms. The means of each form are recorded as the bottom row as μ_f. The population true mean (μ_t) is 1.50. Relevant error variances have been computed from the data as well as estimated from the "Compleat Formula" at the bottom of the table (Refer to Note 4 on Table IV.)

The reliability of the means (Table IV) is less than the reliability of the population (0.33 V.S. 0.60). This is caused by the finiteness of the population! The true variances of the means is a function of $\left(-\frac{n-1}{N-1}\right)$ but the error variance is not. As stated in the Compleat Formula for the Standard Error of the Mean:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_T^2}{n} \left(1 - \frac{n-1}{N-1} \right) + \frac{\sigma_e^2}{n}$$

As noted on Table IV where n=3, the standard error of the mean as determined by the data, is 0.15. Substituting the appropriate data from Table III into the preceding equation yields:

$$\sigma_{\bar{x}}^2 = \frac{0.45}{3} \left(1 - \frac{3-1}{4-1} \right) + \frac{0.30}{3} = 0.05 + 0.10 = 0.15$$

Peters and Van Voorhis (p. 162, formula 89) indicate the correlation of sample means is the same as the correlation for the population. This is now seen to be an incomplete relationship. The compleat formula for the correlation or reliability of sample means can be easily derived. The Compleat Formula for the Standard Error of the Mean defines error variance of the sample means as: $\frac{\sigma_x^2}{n}(1 - r_{xx})$, and variance of the sample mean as:

$$\frac{\sigma^2}{n} \left(1 - r_{xx}\frac{n-1}{N-1}\right)$$

Substituting these into a traditional formula for reliability:

$r_{xx} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$, one arrivies at: $$r_{\bar{x}\bar{x}} = 1 - \frac{\frac{\sigma_x^2}{n} (1 - r_{xx})}{\frac{\sigma_x^2}{n} \left(1 - r_{xx}\frac{n-1}{N-1}\right)}$$

The $\frac{\sigma_x^2}{n}$ cancels leaving:

$$r_{\bar{x}\bar{x}} = 1 - \frac{1-r_{xx}}{1-r_{xx}\frac{n-1}{N-1}}$$

TABLE I

Hypothetical Response Matrix of Four Persons
to
Five Test Items

Person	Item 1	2	3	4	5
A	R				
B		R		R	
C	R	R	R		
D	R	R	R		R

TABLE II

Hypothetical Composition of Four Person Samples
and
Ten Test Samples

(a) Composition of Four Hypothetical Samples of Persons	(b) Composition of Ten Hypothetical Test forms of three items each
Samples	Test Item Form Numbers
A B C	I - 1,2,3
A B D	II - 1,2,4
A C D	III - 1,2,5
B C D	IV - 1,3,4
	V - 1,3,5
	VI - 1,4,5
	VII - 2,3,4
	VIII - 2,3,5
	IX - 2,4,5
	X - 3,4,5

TABLE III

Hypothetical Score Matrix for Four People on each of Ten Test Forms

TEST FORM

PERSON	I	II	III	IV	V	VI	VII	VIII	IX	X	T
A	1	1	1	1	1	1	0	0	0	0	0.6
B	1	2	1	1	0	1	2	1	2	1	1.2
C	3	2	2	2	2	1	2	2	1	1	1.8
D	3	2	3	2	3	2	2	3	2	2	2.4
μ_f	2.00	1.75	1.75	1.50	1.50	1.25	1.50	1.50	1.25	1.00	1.5 μ_t

$$\sigma_x^2 = 0.75$$

$$\sigma_e^2 = 0.30$$

$$\mu_t = 1.50$$

$$r_{xx} = 0.60$$

$$\sigma_T^2 = 0.45$$

TABLE IV

Hypothetical Score Means for Four Samples of Persons and Ten Forms of a Test

Sample	I	II	III	IV	V	VI	VII	VIII	IX	X	$\bar{T}$
A B C	1.67	1.67	1.33	1.33	1.00	1.00	1.33	1.00	1.00	0.67	1.20
A B D	1.67	1.67	1.67	1.33	1.33	1.33	1.33	1.33	1.33	1.00	1.40
A C D	2.33	1.67	2.00	1.67	2.00	1.33	1.33	1.67	1.00	1.00	1.60
B C D	2.33	2.00	2.00	1.67	1.67	1.33	2.00	2.00	1.67	1.33	1.80
μ_f	2.00	1.75	1.75	1.50	1.50	1.25	1.50	1.50	1.25	1.00	1.50 $= \mu_t$

Note 1: μ_f equals true sample means calculated for each test form; $r_{xx} = 0.33$, and $\sigma_f^2 = 0.10$, i.e. the error variance associated with sampling of tests.

Note 2: $\bar{T}$ equals true sample means calculated for each sample of persons; $\sigma_{\bar{T}}^2 = 0.55$, i.e. the error variance associated with sampling of persons.

Note 3: $\mu_t = 1.50$ and represents the true calculated population mean with total error variance $\sigma_{\bar{x}}^2 = 0.15$.

Note 4: The calculated $\sigma_{\bar{x}}^2$ value of 0.15 is also derived by substituting the appropriate data into the "Compleat Formula":[1] Namely:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{nn}\left(1 - r_{xx}\frac{n-1}{N-1}\right) = \frac{0.75}{3}\left(1 - (.60)\frac{3-1}{4-1}\right) = 0.15$$

[1] Donald O. Shepherd and Casimer S. Winiewicz, Ibid.

Substituting appropriate data from Table (3) into the above formula yields:

$$r'_{\bar{x}\bar{x}} = 1 - \frac{1 - 0.60}{1 - (0.60)\frac{3-1}{4-1}} = 0.33$$

The result is exactly as calculated from Table (4). It is easy to see that where N=infinity, the $r_{\bar{x}\bar{x}} = r_{xx}$ It should also be noted that where n=N, then $r_{\bar{x}\bar{x}} = 0.00$ unless $r_{xx} = 1.00$. In that case, $r_{\bar{x}\bar{x}}$ would be indeterminate $(1 - \frac{0}{0})$.

The comparisons presented in Table IV clearly illustrate the usefulness of estimating the standard error of the mean by the "compleat formula" which combines variances associated with both people and tests simultaneously. These results imply that the "compleat formula" should be used in all cases when the population is finite and the reliability is not perfect.

Presented at the MTA Conference
17-21 October 1977
San Antonio, Texas

ABSTRACT

A METHODOLOGY FOR ESTIMATING THE COST-EFFECTIVENESS OF ALTERNATIVE PRETESTS

BY
JACK H. HILLER, Ph. D.
ARMY RESEARCH INSTITUTE FIELD UNIT, PRESIDIO OF MONTEREY, CA

The purpose of the research reported here was to develop a methodology for measuring the cost-effectiveness of alternative pretesting procedures so that an optimal procedure may be selected. The research was accomplished as follows: Variables that affect the amount of time saved or lost by employing pretests were identified and defined. An algebraic model which takes into account measurement accuracy and the affect of pretesting time was constructed so that the amount of time saved (or lost) by pretesting could be estimated. Alternative pretest procedures were formulated. A limited sample of empirical data was gathered to test the cost-effectiveness of the alternative pretest procedures, using a highly efficient data collection procedure. Estimates of the sampling distributions for the variables in the cost-benefit model obtained from the empirical data were used to perform a Monte Carlo study of the cost-benefit values for the alternative pretest procedures.

A METHODOLOGY FOR ESTIMATING THE COST-EFFECTIVENESS OF ALTERNATIVE PRETESTS

BY
JACK H. HILLER, Ph. D
ARMY RESEARCH INSTITUTE FIELD UNIT, PRESIDIO OF MONTEREY, CA

A system for providing on-the-job individual skill training to infantry soldiers is currently being developed by the Army Research Institute under sponsorship of the Army Training Board. Certain key features of this system, which also have relevance to this report, are listed below:

-- Performance-oriented training and testing based on Soldier's Manual tasks

-- Decentralization of individual skill training in which the immediate supervisor is the primary trainer

-- Pretesting to avoid unnecessary training

The primary trainer will often be an infantry squad leader who may have as many as nine soldiers in his squad. The squad leader in his role of trainer is supposed to conduct pretesting to enable more efficient use of training time. However, since the new system requires performance testing as well as performance-oriented training, the degree to which pretesting increases the system's efficient use of time is an unanswered question.

Consider the following hypothetical situation. A squad leader with a nine-man group intends to have all of his men proficient in performing some task. It happens that he has no information concerning any of his men's ability to perform the task, so he conducts a pretest. The pretest chosen is a performance test that may be administered to only one man at a time and requires about ten minutes to conduct for each man. Altogether then, conducting this pretest for the entire squad consumes one and one-half hours of the trainer's time, and during this period the squad members may not be spending their time usefully. If it turned out that every squad member failed the pretest and needed substantial training, then no time was saved, and one and one-half hours were lost.

Given that pretesting may not necessarily yield time savings, the purpose of the research reported here was to develop a methodology for measuring and predicting the cost-effectiveness of alternative pretesting procedures so that an optimal procedure (including exclusion of any pretesting) may be selected for a given situation.

This research effort will be described according to the following stages:

a. Alternative pretesting procedures were constructed.

b. Variables which may affect the time saved or lost by employing the pretesting procedures were identified.

c. A cost-effectiveness model was formulated.

d. An efficient data collection procedure was designed.

e. A computer program to perform Monte Carlo studies with the cost-effectiveness model was written.

Each of these stages is described below.

CONSTRUCTION OF ALTERNATIVE PRETESTING PROCEDURES

The following procedures were considered for possible use as pretests:

a. Squad member's self-estimation of task proficiency;

b. Squad leader's estimations of task proficiency of his squad's members;

c. Paper-pencil criterion-referenced testing;

d. Simulation;

e. Performance testing.

Self estimates are fast and easy to acquire, but of uncertain validity. Given the high levels of turbulence found in operational units, the infrequent occurrence of many infantry tasks, and the fallibility of memory, squad leader estimation was rejected for this research effort. Paper-pencil tests are relatively quick and easy to administer, but have uncertain validity for infantry troops due to their verbal requirements. Simulated testing was considered but rejected due to resource requirements beyond the scope of this project. Hands-on performance testing is typically time consuming, especially for process testing which requires observation of each task step. But performance testing represents the criterion measure for all practical purposes.

While there are clear limitations on each of the above pretests, a decision was made to explore the use of self estimation, paper and pencil criterion- referenced tests, and performance tests.

One possible approach to using the three candidate pretests would be to employ each one by itself. For example, a soldier could be asked if he was able to perform a task to standard, and then either be placed in training if he said "no", or assigned to some other activity if he said "yes". Another possible approach is to arrange the candidate pretests in a systematic order to capitalize on their virtues while minimizing the effects of their weaknesses. Figure 1 shows an ordering of the pretests which may provide an optimal procedure, in terms of its cost-effectiveness. What has been done was to design a procedure which provides the easiest and fastest pretest as the first step, the second easiest to administer as the second step, and the most time consuming pretest, the hands-on performance test, as the last step for anyone who is not already eliminated.

Requiring soldiers to take the performance test as the last step insures that no one will falsely be considered proficient. Theoretically, the only error that can be made with the above ordering of pretests is to assign a soldier to training when he does not need it. Such errors may occur either because the soldier misjudges his true ability, or because he fails the paper-pencil test.

In addition to the pretesting procedure shown in Figure 1, several other possible procedures were defined for this research. These were:

a. Self estimate followed by performance test;

b. Paper-pencil test followed by performance test;

c. Performance test alone;

d. No pretesting, everyone enters training.

These procedures are diagramed in Figures 2-4.

COST-EFFECTIVENESS VARIABLES

Having defined alternative pretesting procedures, it is necessary to devise a method for identifying an optimal pretesting procedure for various training situations. The approach taken was to identify variables which may influence time saved or lost, and then to construct an algebraic model which may be used to calculate the time saved or lost. Key variables that were identified are:

a. The proportion of group members who are proficient before training is conducted.

b. The proportion of these proficient group members who are correctly identified by a pretest, and the proportion who are incorrectly classified as needing training.

c. The time it takes to conduct the pretest procedure.

d. The time it takes to conduct the mandatory end-of-training performance test.

e. The time it takes to conduct the first phase of training--the explanation/demonstration--after which a trainee may request to be given the post-training performance test.

f. The time benefit (or loss) accruing from pretesting, which is defined as the total number of man-minutes saved by pretesting.

The benefit equation is too involved to explain in the limited time available here, but it is presented and explained in the handouts for two kinds of pretesting errors, that is, assigning a soldier to training even though he is proficient, or failing to assign a soldier to training despite the fact that he is not proficient.

AN EFFICIENT DATA COLLECTION PROCEDURE

The alternative pretesting procedures described above need to be comparatively evaluated from the standpoint of their cost-effectiveness. To do this, a domain of tasks for which training would be conducted must first be specified, as for example the 57 common or basic tasks included in the Soldier's Manual for MOS 11B. Given the fact that a data collection effort involving the four different pretesting procedures would be time consuming and difficult to accomplish, it seemed important to devise an efficient data collection procedure. The solution devised was as follows. Participating soldiers are asked to read a description of the task, conditions, and standards for each task that is sampled from the Soldier's Manual. They are then asked first to estimate their ability to perform it, second to take a paper-pencil test regardless of their estimate, and finally to take the performance test regardless of their written test result. The data collected by this procedure may then be distributed to all four active pretesting alternatives by means of a logic tree analysis, thereby effecting a sizeable economy in data collection requirements. The analytic technique is illustrated by the results for a sample task in Figures 5-8.

MONTE CARLO PROGRAMMING

Data have been collected from a small sample of soldiers for only four Soldier's Manual tasks. The difficulties encountered when collecting this data from troops in the field motivated planning to take maximum advantage of such data. To this end, a Monte Carlo computer program was written to simulate the sampling distributions of predicted benefits for the alternative pretesting procedures. The means and variances for each variable in the benefit equation were estimated from the sample of

soldiers who were pretested on the Soldier's Manual tasks. These parameters were used as input to the Monte Carlo program. The program then generates an estimate of the sampling distribution for the benefit values of each of the alternative pretesting procedures. A one way ANOVA may then be used to a test for significant differences among the mean benefit values associated with each of the alternative procedures.

RESEARCH PLANS

Current plans call for collection of data from two infantry battalions using a sample of 16 Soldiers' Manual tasks. Apart from any specific results which may be obtained from this sample, the methodology described in this paper may prove useful in other applications where an optimal cost-effective procedure needs to be selected from a set of alternatives.

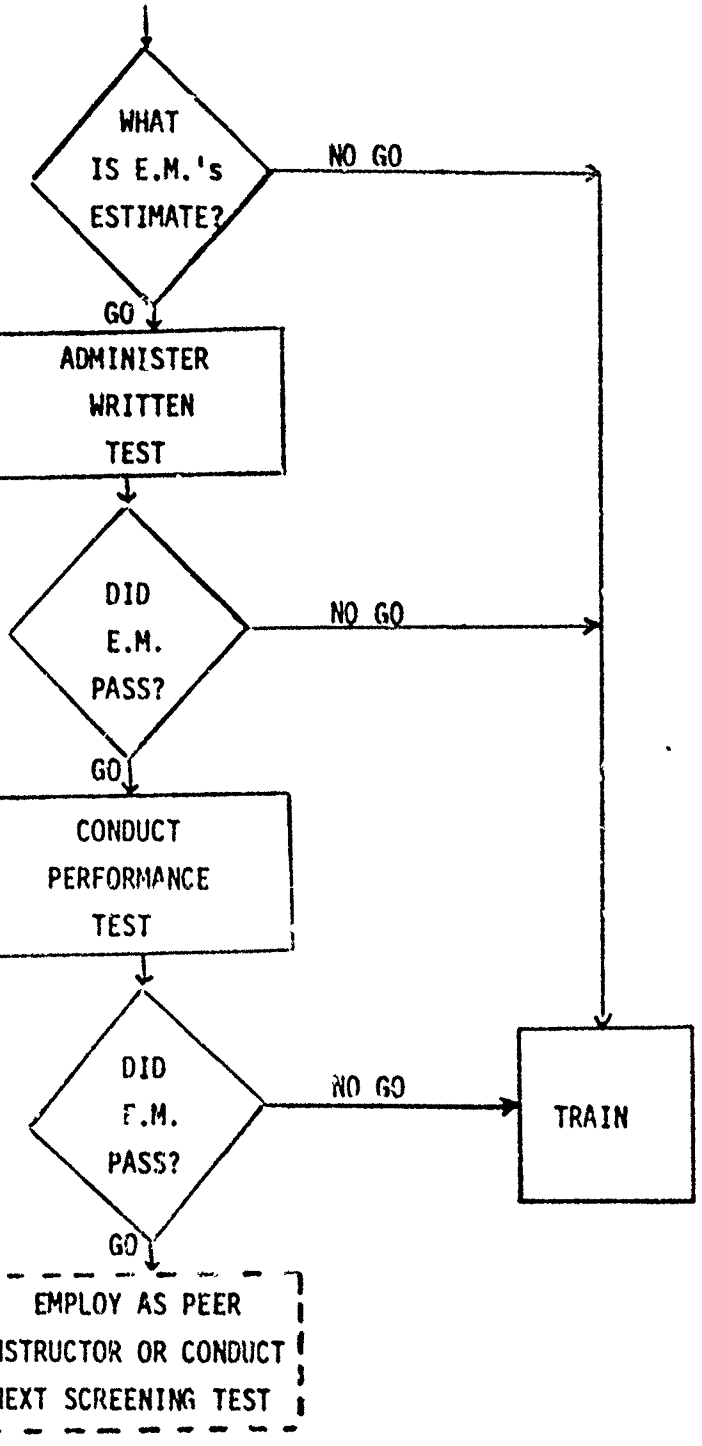


Figure 1. Pretesting Model A.

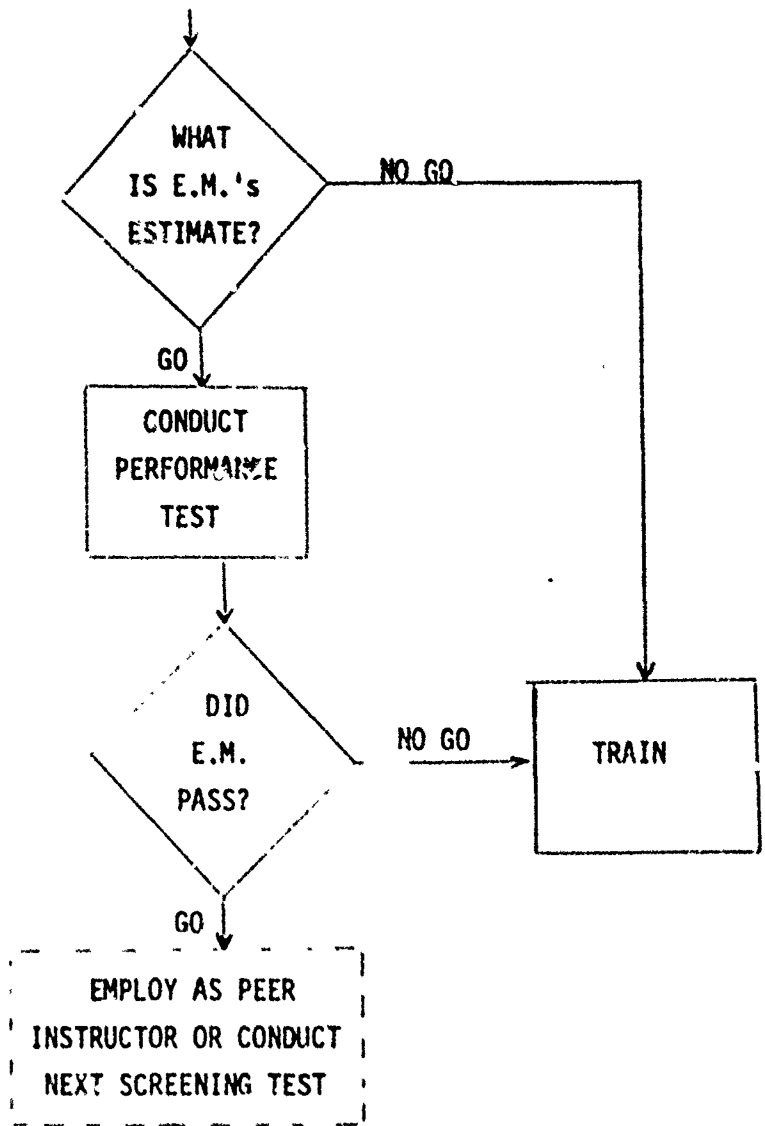


Figure 2. Pretesting Model B.

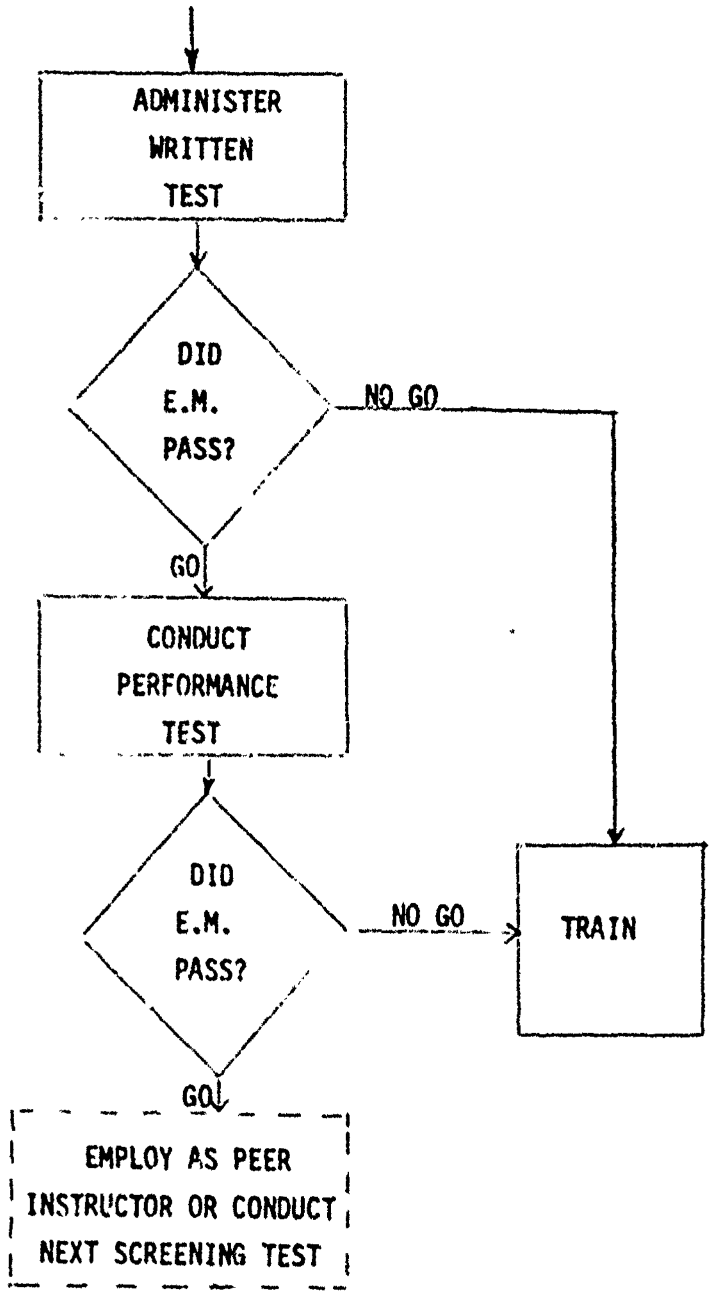


Figure 3. Pretesting Model C.

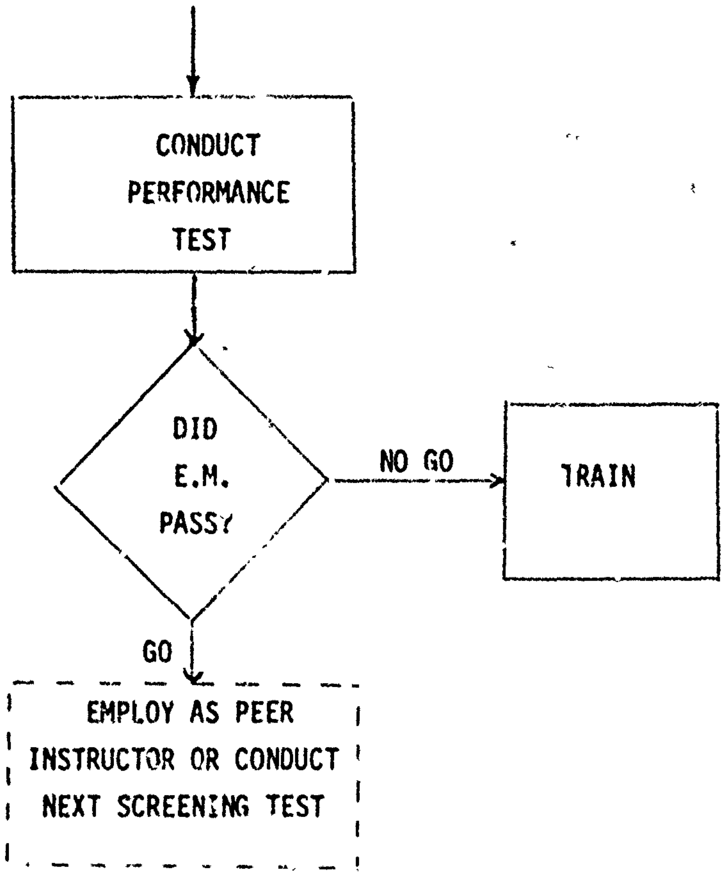


Figure 4. Pretesting Model D.

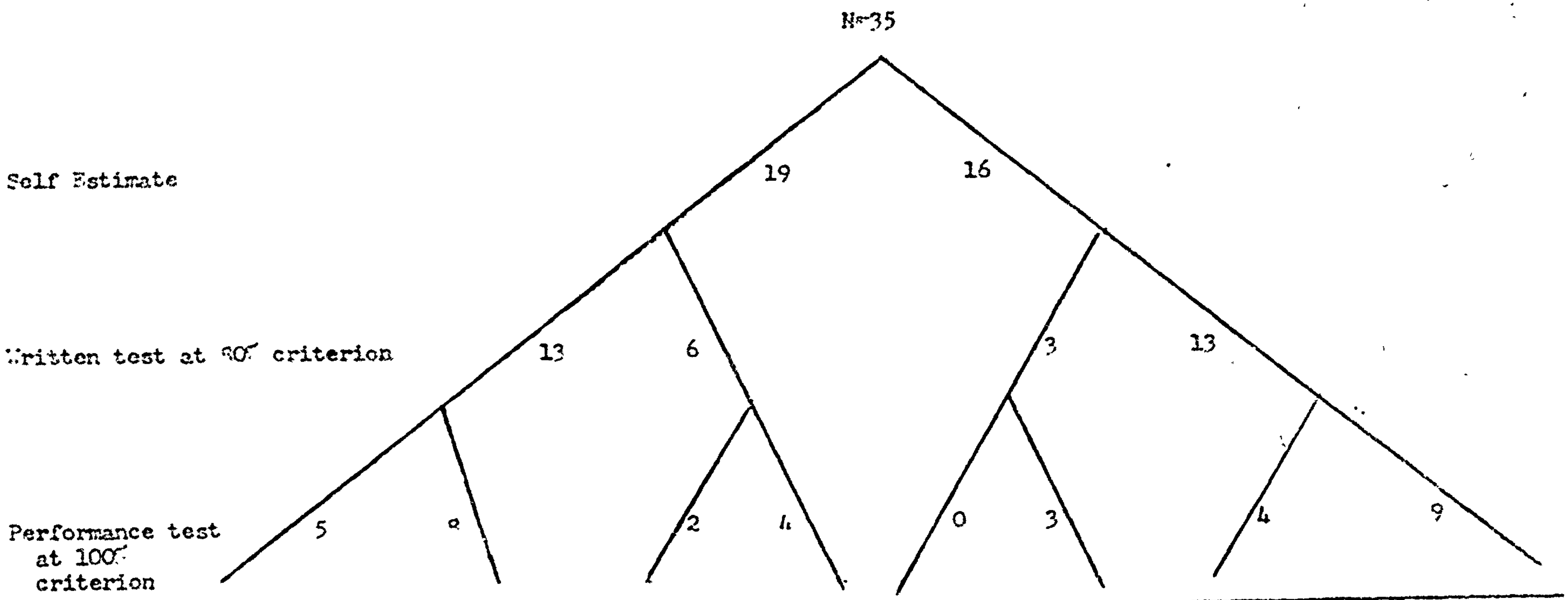


Fig. 5. Data collected for the Soldier's Manual Task -- Encode/Decode KAL61. The paths to the reader's left(/) are used for Go results and to the right(\) for No Go results. Thus, 5 soldiers passed all pretests and 9 passed none.

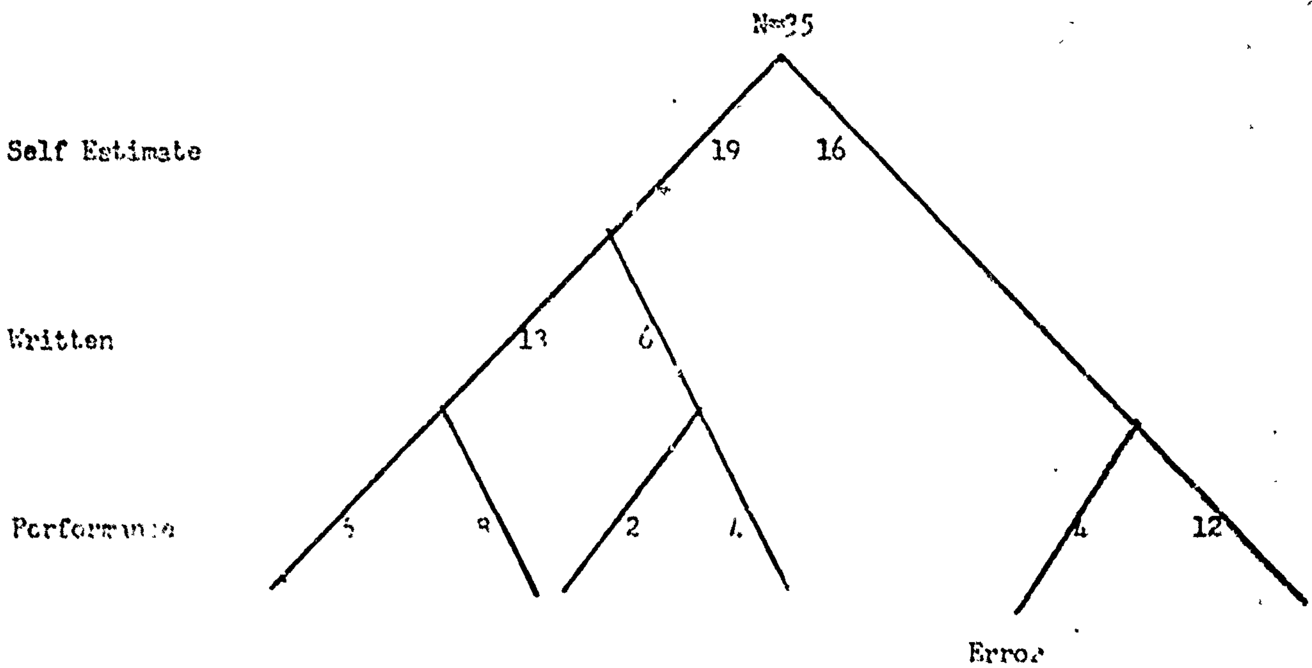


Fig. . Results for task -- Encode/Decode KAL61 -- needed to calculate the benefit value for Model A.

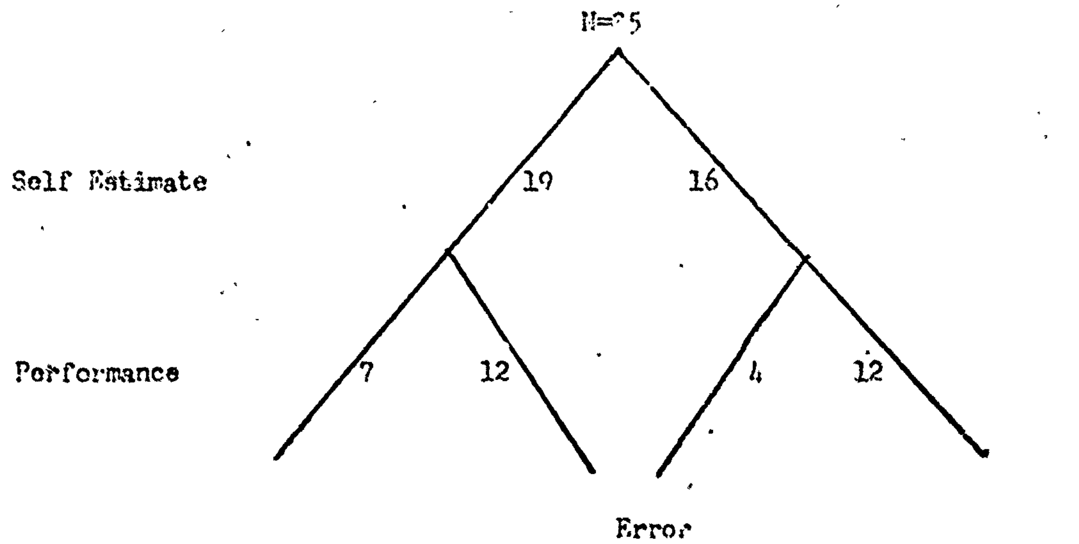


Fig. 7. Results for task -- Encode/Decode KAI61 -- needed to calculate the benefit value for Model B.

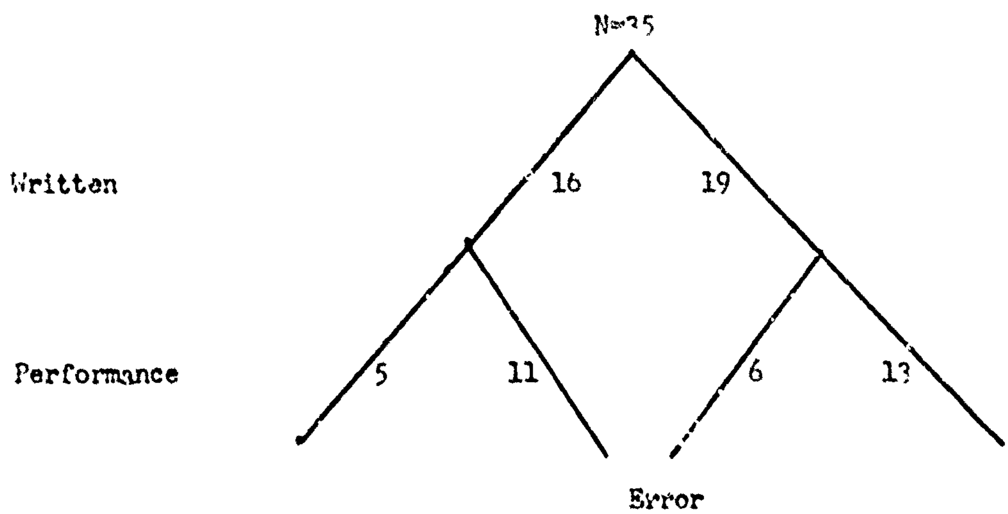


Fig. 8. Results for task -- Encode/Decode KAI61 -- needed to calulate the benefit value for Model C.

A Benefit Model Applying Where "GOs" May Be Misclassified "NO GO"

Listed below are the variables used by a cost-benefit model in which the only measurement error is classifying a task proficient, or "GO," individual as "NO GO":

B = Benefit defined as the total number of man-minutes saved by applying any pretesting procedure.

G = The proportion of squad members who are able to perform the task (i.e., who are "GO") before training is given.

E_N = The proportion of squad members who are able to perform the task but are erroneously classified by a pretesting procedure as "NO GO."

$G - E_N$ = The proportion of men in a squad who are correctly classified by a pretesting procedure as "GO" on a task.

P = The time that it takes a squad leader to conduct a pretesting procedure for his entire squad.

C = The time it takes to give a performance test or checkout to one man; C becomes zero when no checkout is given a pretest.

D = Time to demonstrate how to perform a task.

N = The number of men in a given squad.

The term $N(G-E_N)D$ yields the amount of time that is saved by not "training" men who are able to perform the task without training. This term yields the primary time savings from any pretesting procedure.

The term NE_ND yields the amount of time that is wasted by "training" men who could perform the task without additional training.

The term NP is the time cost incurred by pretesting. The term $N[G-E_N]^2C$ is subtracted from P to provide a credit for the time that would have been spent in conducting a checkout on qualified performers after training, as required by the IETS model.

The benefit model is formed by algebraically combining the above terms as follows:

$$B = N(G-E_N)D - NE_ND - N(P-N(G-E_N)^2C)$$

$$= N[D(G-2E_N) - P + N\ (G-E_N)^2C]$$

A Model for Estimating Time Lost When "NO GOs" May Be Misclassified "GO"

Presented next are the variables and a model for computing the time lost by the squad leader when he has to conduct a special training session for individuals whom he had previously incorrectly classified as "GO."

E_G = proportion of squad members erroneously classified as "GO" who are actually "NO GO."

R = time lost by squad leader by having to provide training to the NE_G group after providing training to the others, $N(1-E_G)$.

O = time it takes squad leader to <u>organize</u> the training (i.e., obtain training materials, move to a training location and set up the training session) for the NE_G group that is in addition to the time spent preparing training for the $N(1-E_G)$ group.

M = time required by an average soldier for supervised practice until he <u>masters</u> the task.

$R = O + D + M$

L = time in man minutes that the squad leader <u>loses</u> when he has to conduct training for the NE_G group after he has already taught the $N(1-E_G)$ group. The group that is losing the squad leader's time is $N(1-E_G)$.

$L = N(1-E_G)R = N(1-E_G)(Q+D+M)$

A Model for Estimating the Benefit from Pretesting Where Both Kinds of Error May Occur

Finally, a model to compute the benefit attributable to pretesting where both kinds of classification error can occur (i.e., "GO" misclassified "NO GO" and "NO GO" misclassified "GO") is presented below.

$$B = N[D(G-2E_N) - P + N(G-E_N)^2C - (1-E_G)(Q+D+M)]$$

Development of the Armed Services Vocational
Aptitude Battery

Malcolm James Ree
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

I. INTRODUCTION

In February of 1966, a joint services committee of measurement and evaluation experts was formed and given the responsibility for the development and standardization of a differential aptitude battery for use in a joint services high school testing program. The primary goal in the development of the battery was to design a single aptitude measurement instrument which would provide adequate coverage of the content included in the classification batteries used by each of the individual armed services.

II. ARMED SERVICES VOCATIONAL APTITUDE BATTERY (ASVAB)

The ASVAB is composed of aptitude measures reflecting the content of the classification batteries used by the Army, Navy, and Air Force and to one which is used in a joint services high school testing program. Accordingly, the Army, Navy, and Air Force batteries were administered to a random sample of 3,900 military basic trainees (Bayroff & Fuchs, 1970). A counterbalanced order of administration was used to prevent possible practice effects. Intercorrelations for all test variables were computed and served as the basis for the selection of aptitude measures common to all three classification batteries. On the basis of these

analyses, nine subtests were chosen and organized into a battery, the Armed Services Vocational Aptitude Battery.

Eight of the nine ASVAB subtests were selected from the Army, Navy, and Air Force batteries. The ninth subtest was a modification of the Army Coding Speed Test. The criteria for item selection in each subtest were mean item difficulty level, a lower limit of acceptance in terms of item discrimination level, and content validity. The items for each of the nine subtests were arranged in ascending order of difficulty within each subtest.

In September of 1968, ASVAB-1 was accepted for use in the High School Military Testing Program. During that same year, Vitola and Alley (1968) developed Air Force aptitude indexes for use in the operational selection and classification program.

In early 1974, the Department of Defense directed that the services move expeditiously toward the use of a common aptitude battery for enlistment qualification. The Office of the Assistant Secretary of Defense (Manpower and Reserve Affairs) suggested that the Armed Services Vocational Aptitude Batteries be redesigned to satisfy enlistment production requirements of all the services with high school usage being a secondary consideration.

Since the introduction of ASVAB-1, several alternate forms have been developed. ASVAB-1 was initially used in the high school testing program and was subsequently replaced by ASVAB-2. In September 1973, ASVAB-3 supplanted the Airman's Qualifying Exam Form J (AQE-J) in the Air Force Airman Selection and Classification Program. ASVAB-4 was essentially a back-up instrument for use in case of test compromise.

The contents of ASVAB Forms 5, 6, and 7 represent a substantial departure from previous forms. The redesign of the ASVAB was based upon the content of the then current armed service classification batteries, the uses being made of this content, and the services' future plans for the modified battery. A preliminary battery plan was developed at the Air Force Human Resources Laboratory (AFHRL) for review at the other service laboratories.

The initial plan called for two perceptual tests, 12 cognitive power tests, and a rather lengthy interest inventory culled from materials from the Army Classification Inventory (ACI), the Navy Vocational Interest Inventory (NVII), and the Air Force's Vocational Interest Choice Examination (VOICE). It was estimated that the battery defined in the initial plan would require over four hours of testing time.

Table 1 presents the contents as proposed in the initial plan.

Table 1. Preliminary Plan for ASVAB[a]

Content Area		Number of Items
Attention to Detail		30
Numerical Operations		50
Word Knowledge		25
Arithmetic Reasoning		25
Space Perception		25
Mathematics Knowledge		25
Electronics Information		25
Radio Information		15
Mechanical Comprehension		25
Automotive Information		25
Shop Information		25
Biological Science		15
Physical Science		15
General Information		20
Interest Inventory		527[b]
Army Classification Inventory	(87)	
Navy Vocational Interest Inventory	(190)	
Vocational Interest Choice Examination	(250)	
	Total	872

[a]Estimated testing time - 4 hours 6 minutes.

[b]Total item pool before consolidation, where possible, of content from the three source inventories.

Table 2 shows the final content of the battery. It was essential that the battery be shortened from its estimated 4 hours and 6 minutes, especially for application in the joint services high school testing program. In addition, the various recruiting service commanders desired a testing time considerably shorter than required in the preliminary plan. The content shown in Table 2 was arrived at after a series of joint service committee deliberations, and was possible only because of various compromises from what would have been considered optimal by each service. Note that with the exception of Word Knowledge, all the scales in the final plan were shortened from the originally planned number of items. This was because the U.S. Coast Guard used Word Knowledge to screen personnel for officer programs and believed that anything less than 30 items would be inadequate for that purpose.

Table 2. Final Plan for ASVAB

Content Area	AFQT[d]	Number of items	Test Time (in minutes)
Attention to Detail		30	5
Numerical Operations		50	3
Word Knowledge	x	30	10
Arithmetic Reasoning	x	20	20
Space Perception	x	20	12
Mathematics Knowledge		20	20
Electronics Information[a]		30	15
Automotive Information		20	10
Shop Information		20	8
General Science[b]		20	10
General Information		15	7
Classification Inventory[c]		87	20
	Total	362	2 hrs 35 min

[a]Composed of 15 Electronics Information and 15 Radio Information items.
[b]Composed of 10 Biological and 10 Physical Science items.
[c]Army items only.
[d]Armed Forces Qualification Test.

ASVAB-5 is a high school version, and ASVAB-6 and 7 are the current operational production tests. Jensen, Massey, and Valentine (1976) may be consulted for a more detailed description of the scales and the relevant normative studies.

III. DEVELOPMENT OF ASVAB FORMS 8, 9, AND 10

The following steps were executed in the construction of ASVAB-8, 9, and 10.

Initial Item Selection and Editing

Approximately 2,500 item cards were culled from the Air Force Human Resource Laboratory historical files. After review, revision, and editing to insure appropriate wording and grammatical agreement of item stems and item options, 2,400 items were selected. Most were "new" items, but a few had been used in previous ASVAB forms or in previous military selection batteries.

Construction of Tryout Booklets

These items were assembled into 16 tryout tests of approximately equal difficulty. Two forms (A and B) of each tryout booklet were prepared so that no item would be always last and thus frequently omitted. The testing time for each booklet was 90 minutes.

Administration of Experimental Items

According to a geographic sampling plan, these 16 tryout booklets were administered at the 64 Armed Forces Entrance and Examination Stations (AFEESs), with each station testing only four booklets. Booklets, answer sheets, and administrative instructions were provided. The

subjects for the tryout cycle were randomly sampled applicants appearing at the AFEESs for possible qualification for military enlistment. The number of subjects required from the various AFEESs ranged from 28 to 84, for a total projected sample of 3,200. Three weeks were allowed for this testing cycle, and answer sheets for 2,588 applicants were received, representing a loss of about 19 percent. No systematic bias was found in the returned answer sheets, and the sample was acceptable for tryout use.

<u>Assembly and Administration of the Proposed ASVAB Forms</u>

Using previous ASVAB (7) items and subscales as models, three new ASVAB forms were developed from items tried out in the experimental testing cycle. The three new forms closely resembled each other and previous ASVABs in terms of individual item difficulty and discrimination level. The testing time for each experimental ASVAB (8, 9, & 10) form was 2 hours and 30 minutes.

The proposed ASVAB (8, 9. & 10) forms were administered at the 64 AFEESs. Only one form was sent to each station for experimental administration to minimize potential compromise problems, and each subject also took the current ASVAB (7). Subjects for this testing cycle were randomly selected from the applicants for military enlistment. The numbers required from each AFEES varied from 52 to 120, for a total projected sample of 6,000.

Five weeks were allowed for this testing, and a total of 4,308 usable sets of answer sheets were returned, representing a loss of about 28 percent. Analysis indicated that the loss did not represent any particular geographic or ability bias. These responses were then used to develop various information about the test: reliability, average item difficulty,

and rough norms. A detailed analysis of ASVAB construction may be found in Fruchter and Ree (1977).

IV. PROVIDING ASVAB NORMS

The studies to provide normative conversions and tables for the ASVAB can be divided into those for high school norms and those for armed services norms.

ASVAB-5 was standardized on a high school sample of 35,291 male and female students in grades 9 through 12. The sample was stratified by geographic area, school size, and the percentage of minority enrollment. Student scores were then weighted to make the sample represent the national high school population. This study produced normative tables by grade and sex for the ASVAB subtests and high school composites. This enabled the ASVAB to be used as a high school guidance tool as well as for military enlistment. A detailed description of the development of high school norms may be found elsewhere (Adkins, 1976).

Jensen, Massey, and Valentine (1976) reported the development of the armed services' norms for ASVAB-5, 6, and 7. Procedures for producing the normative data included the administration of ASVAB-5, 6, and 7 to a nationally representative sample of applicants for military enlistment at the 64 AFEESs. Examinees took one form of the ASVAB (5, 6, or 7) and either the Armed Forces Qualification Test (AFQT) composite from the Army Classification Battery or the ASVAB-3 in a counterbalanced administration. From the responses of approximately 4,500 examinees evenly

divided among ASVAB-5, 6, and 7, a stratified sample of 1,600 responses was used to compute percentile equivalents for each raw score value on all subtests and military composites.

V. STANDARDIZATION AND EQUATING OF ASVAB FORMS

A sample of 2,052 male and female high school students in grades 9 through 12 from 26 schools participated in a study to standardize the ASVAB-5 to ASVAB-2. Test administration conditions were uniform within each school for the presentation of Form 2 and Form 5 on two consecutive days. For reasons of maximum comparability (Flanagan, 1951), tests were scheduled at the same time and in the same rooms on either two consecutive mornings or two consecutive afternoons. The tests were administered in a counterbalanced design with an equal number of subjects having the tests administered in the order ASVAB-2 then ASVAB-5, and the order ASVAB-5 then ASVAB-2.

The scores were equated by an equipercentile method (Angoff, 1971). Then, using a technique originally implemented by Lindsay and Prichard (1971), the fitting of the curve to the equated data points was done by an interative least-squares regression procedure.

The results of this study were a series of tables equating scores of subtests and composites on ASVAB-5 to percentile scores of subtests and composites on ASVAB-2. These conversion tables were developed for each grade, for males, for females, and for both sexes (Fletcher & Ree, 1976).

ASVAB scores have also been equated to scores on the Differential

Aptitude Test Battery (DATB) and the General Aptitude Test Battery (GATB) (Kettner, 1976). From two schools, 1,232 ninth through twelfth grade males and females were tested on either the ASVAB and DATB or on the ASVAB and GATB. The tests were administered in a counterbalanced order and then scored. Equations for predicting ASVAB subtests from DATB subtests and for predicting ASVAB subtests from GATB subtests were developed. These scores on the DATB or GATB were then equated to ASVAB scores. This helped in understanding the nature of all the tests involved and provided the counselor with another tool for vocational assessment.

VI. FUTURE PLANS

The development of high school and operational forms of ASVAB will continue. In addition, the development of adaptive testing technology will be studied and monitored. Already three subtests of ASVAB, Word Knowledge, Arithmetic Reasoning, and Space Perception (the AFQT) have been adapted for presentation via computer media (Ree, 1977). As this area of technology develops, its application to meeting the goals of ASVAB testing will be explored.

REFERENCES

Adkins, E. Armed Services Vocational Aptitude Battery (ASVAB) Form 5: High school norming and standardization (AFHRL-TR-76-79). Brooks AFB TX: 1976.

Angoff, W. Scales, norms, and equivalent scores. In Thorndike, R. (Ed.) Educational Measurement (2nd ed.). Washington DC: American Council on Education, Chapter 15, 1971.

Bayroff, A., & Fuchs, E. The Armed Services Vocational Aptitude Battery (Technical Research Report 1161). Arlington VA: U.S. Army, 1970.

Flanagan, J. Units, scores, and norms. In Lindquist, E. (Ed.) Educational Measurement. Washington DC: American Council on Education, Chapter 17, 1951.

Fletcher, J., & Ree, M. Armed Services Vocational Aptitude Battery (ASVAB) correlational analysis, ASVAB Form 2 versus ASVAB Form 5 (AFHRL-TR-76-70). Brooks AFB TX: 1976.

Fruchter, D., & Ree, M. Development of the Armed Services Vocational Aptitude Battery: Forms 8, 9, and 10 (AFHRL-TR-77-19). Brooks AFB TX: 1977.

Jensen, H., Massey, I., & Valentine, L. Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7) (AFHRL-TR-76-87).

Kettner, N. Armed Services Vocational Aptitude Battery (ASVAB Form 5): Comparison with GATB and DAT tests (AFHRL-TR-76-78). Brooks AFB TX: 1976.

Lindsay, C., & Prichard, M. An analytical procedure for the equipercentile method of equating tests. Journal of Educational Measurement, 1971, 8, 203-207.

Ree, M.J.. Implementation of a model adaptive testing system at an Armed Forces Entrance and Examination Station. An invited paper for "Adaptive Testing '77," Minneapolis, July 1977.

Vitola, B., & Alley, W. Development and standardization of Air Force composites for the Armed Services Vocational Aptitude Battery (AFHRL-TR-68-110). Brooks AFB TX: 1968.

PREDICTIVE VALIDATION OF ASVAB FORMS 6/7
FOR
NAVY ENTRY LEVEL TECHNICAL TRAINING COURSES

Leonard Swanson

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

Problem

The Asmed Services Vocational Aptitude Battery (ASVAB), Forms 6 and 7 have been used for selection and initial assignment of recruits by all Armed Services under the Department of Defense since January 1976. This Battery, which contains tests similar to those in the earlier classification batteries of the various services, has been validated in only a small portion of Navy schools. In order to maintain effective standards for Class A-school selection based on ASVAB tests, the Bureau of Naval Personnel has requested additional evaluation of the ASVAB for predicting performance in Class A-schools. A recent increasingly important problem has been the unacceptably high attrition from Basic Electricity and Electronics (BE&E) School. A special effort to reduce attrition by changes in selector tests was made for a set of electromechanical ratings in conjunction with the corresponding A-schools.

APPROACH

Samples

The ASVAB was administered by classification testing personnel to Navy applicants at the time of enlistment at an Armed Forces Entrance and Examining Station, at a mobile examining test site or at a Naval Training Center. Subsequently, most of the accepted applicants were assigned to various Navy Basic and Class A-Schools for training. Forty-one A-Schools were included in this validity study with sample sizes presented in Table 1. Students in most of included schools completed school training by December 1976. Students in a few schools completed school training as late as April 1977. Most of these samples, do not include all or even a majority of students who completed school during 1976 for various reasons: (1) students beginning school training before April 1976 had entered the Navy when the Basic Test Battery rather than the ASVAB was used for classification, (2) school criterion data were not available by the cut off date for inclusion in the sample sets, and (3) the number of academic drops for some schools is smaller than is known to be the case.

Variables

1. Predictors. All the separate ASVAB tests variables are reported as Navy Standard Scores (NSS) having a mean of about 50 and a standard deviation of 10 for an unrestricted recruit population. The 12 subtests are:

General Information (GT): A 15 item test of general knowledge which includes questions on sports, outdoor activities, automobile mechanics, and history. Testing time is 7 minutes.

Numerical Operations (NO): Measures how rapidly and accurately the examinee can add, subtract, multiply, and divide small whole numbers. Testing time is 3 minutes for 50 items.

Attention to Detail (AD). This tests a person's ability to pick out details rapidly. Each item contains two lines of c's and o's. The number of c's must be counted and the correct answer selected from five alternatives. Testing time is 5 minutes for the 30 items.

Word knowledge (WK): This test presents 30 vocabulary words. The examinee must select from four alternatives the word which most nearly has the same meaning as the given word. Testing time is 10 minutes.

Arithmetic Reasoning (AR). This test consists of 20 reasoning problems in sentence form. The examinee must solve each problem and select the correct answer from four alternatives. Testing time is 20 minutes.

Space Perception (SP): A 20 item pictorial test consisting of flat patterns and drawings of three-dimensional geometrical figures. Broken lines on the flat pattern show where it is to be folded. The examinee's task is to select the three-dimensional figure which could be made from the flat pattern or to select the flat pattern which represents the three-dimensional figure. Each item has four alternatives. Testing time is 12 minutes.

Mathematics Knowledge (MK): A 20 item test which requires some knowledge of algebra, geometry, fractions, decimals, and exponents. The correct answer must be selected from four alternatives. Testing time is 20 minutes.

Electronics Information (EI): A 30 item test of the examinee's knowledge of electrical and electronic components, principles, symbols, and diagrams. The correct answer must be selected from four alternatives. Testing time is 15 minutes.

Mechanical Comprehension (MC): In this 20 item test a drawing illustrates a mechanical principle and a question is asked about the drawing. The correct answer must be selected from four alternatives. Testing time is 15 minutes.

General Science (GS): This is a 20 item test of knowledge of physical and biological science. Each item has four alternatives. Testing time is 8 minutes.

Shop Information (SI): This test consists of 20 questions about shop practices and the use of tools. Some of these four alternative questions are pictorial. Testing time is 8 minutes.

Automotive Information (AI): This test has 20 questions about automobile parts and their operations. Each item has four alternatives. Testing time is 10 minutes.

Along with the twelve individual cognitive ASVAB tests sixty-three combinations of tests were included as predictors. These included the four commonly used Navy composites and five special composites for indiviudal Navy schools anf fify-four other combinations of ASVAB tests to discover alternate composites that might prove to be more valid than existing ones. Of these sets, twenty-three were 2-test sets, twenty-five were 3-test sets and six were 4-test sets. The compoistes ued by the other services were included as variables. Several

of the Army composites include one of the four scales in the Classification Inventory (CI), the thirteenth test in the ASVAB. The Classification Inventory scales were not included in the Army composites because the CI scale scores were not included on the Navy Enlisted Master tape extract, the source of ASVAB scores.

2. Criteria. Class A-School criteria were obtained from individual schools on a school reporting from provided to the Navy Personnel Research and Development Center (NPRDC). Final School Grades (FSG) was available for nineteen A-Schools. For the other twenty-two A-Schools using a self-paced mode of instruction, a Days-in-Training (DAYS) criterion was used. This was computed from the course starting and completion dates reported for the students. A pass-fail criterion was obtained for BE/E school students from the Chief of Naval Education and Training.

Data Analysis

Means, standard deviations and correlations among predictors and A-School criterion variables were computed for each school. The validity for each predictor was corrected for restriction in range which occurred when students were selected for technical training.[1]

Multiple correlations were computed from uncorrected correlations for the twelve ASVAB cognitive tests for each course. An accretion method was used in which a multiple correlation was computed after the addition of each test.

RESULTS AND DISCUSSION

The basic validity data for each of the forty-one A-Schools are presented in Table 2. The schools are arranged in alphabetic order within selector composite groups, with schools having a Final School Grade criterion listed first followed by those with a Days-in-Training criterion. Uncorrected and corrected zero-order validites are presented as well as uncorrected multiple correlations for all twelve ASVAB cognitive tests and for the most valid set of five and three tests in each school. Validities of current selector composites and of tests included in them are underlined.

From Table 2 it can be seen that ASVAB validity depends substantially on what criterion of school performance was used. For the nineteen A-schools which have a Final School Grade criterion and median uncorrected validity of the current selector composite is .46, with a range from .16 to .69. The

[1]Since A-school students must have minimum classification test scores usually above the mean, test validities obtained for school samples are lower than would be found for a sample with a broader range of ability. The obtained validities can be adjusted or "corrected" to reflect what the validities would be for a sample covering a full range of ability. This also permits a fairer comparison of test validities for schools with different required scores on classification tests. The formula used for corrected correlations are presented in Guilford, J. P. Fundamental Statistics in Psychology and Education, New York: McGraw Hill, 1956, p320-321. Case I was used to correct the validity of the variable used in selection. Case III was used for other variables.

median corrected validity is .64, indicating substantial predictiveness about equal to that previously reported for the Navy Bast Test Battery (Thomas, 1970).

Twenty-two of the schools in this analysis are self-paced and do not compute a Final School Grade for their students. In these schools the course is customarily divided into modules of instruction. The student must pass a test on each module with a minimum grade of 90% before advancing to the next module. For these schools the median validity of the ASVAB selector composite against a Days-in-Training criterion is only -.075, with a range from .18 to -.29. (A negative validity is expected for the Days-in-Training criterion since fewer days to complete a course should reflect greater ability.)

The picture is not completely dismal for the Days-in-Training criterion. For the nine self-paced schools using the Mechanical or Electronics composites the median uncorrected validity of the current selector composite is -.21. The median corrected validity is -.39. While these values are less than satisfactory they are not zero as is the case for the thirteen self-paced schools using the General Technical or Clerical selector composites.

We do not yet fully understand why the validities for self-paced courses are so low. We know of some factors that could reduce the validities, but do no know to what extent these factors are present. For example, in discussions with school administrative personnel it was learned that some students could have finished the course earlier than they did, but postponed completion of the course untill the end of a week rather than finish early and be assigned to General Detail for a few days while awaiting transfer to a new duty station. In some self-paced schools the variance of Days-in-Training is small, not much more than it is in some lock-step courses.

This is a problem that has not yet received as much attention as it deserves. At the last MTA meeting Dr. Raymond Christal had some very worthwhile comments related to this topic as well as suggestions for research (Christal, 1976). It is an especially important problem because more and more schools have gone from a lock-step mode of instruction to a self-paced mode over the past few years and the trend is continuing.

The maximum validity of the ASVAB is whown for each school by the multiple correlation using all twelve ASVAB tests. The median R across the 19 schools with FSG is .60, somewhat higher than the .46 I mentioned earlier for the median uncorrected validity of the current selector composite. There seems to be room for some increase in validity. Also shown are the validities of the most valid set of five and three ASVAB tests. There is very little difference between validity of the five most valid tests and all twelve ASVAB tests and, for some schools, not much difference between the most valid set of three and all twelve tests. The reason for listing validities of five and three tests is that no current ASVAB composite has more than five tests, and three is a more usual number of tests comprising a composite.

From examining the validities of ASVAB subtests and selector composites you can see that the operational composites do not have the highest validites for many schools. Zero order validities of over fifty sets of 2-, 3-, or 4-ASVAB tests were compared with the current selector composites. This inspection showed that seldom did the 4-test linear sums yield higher validities than the best 2- and 3-test sets. Therefore, the most valid sets of 2- and 3-tests were extracted and are shown in Table 3 along with the validities of the operational

selector composites for each school. It can be seen that for most schools several, or sometimes many, sets of tests yield almost the same validities. This reflects the high relationships among many of the ASVAB tests. These data were examined, along with similar data from a concurrent ASVAB validity report on thirty-one schools (Swanson, 1976) and, in some cases, earlier Navy Basic Test Battery validity data (Thomas, 1970, 1973) on the same schools in order to arrive recommended changes in selector composites.

The Basic Electricity and Electronics (BE/E) course is a prerequisite for over twenty A-schools. There has been an excessive attrition rate for BE/E students destined for some of these ratings. Of particular concern were four electromechanical ratings—EM, IC, CE, and GM. A study by Dann and Abrahams (1977) examined attrition in BE/E School, and considered these four ratings. The present study evaluates the joint impact of selector composite changes on BE/E and "A" school. There are Class A-school samples in this study for only two of these four ratings—EM and GM. Analyses of ASVAB test validities were made for the separate groups of BE/E students destined for the individual schools and for BE/E students with a common selector composite. Regression analyses was used with hold out samples for cross-validation. The details of the procedures and results are presented in an unpublished report by Dann and Abrahams (1977). The validities of the current selector composites and the most promising new selector composites with integer weights for each test rather than the more precise regression weights for the four school samples are shown in Table 4. The validities of the recommended selector composites are enclosed in a dotted box. For Gunner's Mate students the recommended Electronics composite yields an increase in validity from .17 to .37 in the BE/E school, where substantial attrition has occurred. This is accompanied by a slight decrease in A-school validity from .46 to .41. The overall effect is expected to be a larger throughput of trained students into the Gunner's Mate rating. For EM students the newly recommended BE/E composite, 2MK+AR+GS, improves BE/E school validity from .11 to .41, which should reduce BE/E school attrition without reducing the validity obtained with the present A-school composite. There is no IC A-school in this analysis. Nevertheless, a change to the new BE/E selector composite for both IC Class A and BE/E students seems warranted on the basis of similarity in course content, job knowledge required and a history of similar validities for like tests for these courses.

The Construction Electrician sample size is too small to make a strong case for change. Validity data for additional students in these and other BE/E and A-schools will be available within a short time to check on these recommendations.

In a least 12 other schools the validities of alternate selector composites were sufficiently higher than validities of the present composites to be considered for operational use. In some of these schools the sample sizes were too small to have confidence that the alternate composite validities would hold up. The schools in which changes in selector composites appear warranted are shown in Table 5. In two schools, MM and EN, validity differences from present composite validities are small. These schools are nevertheless included in order to have a uniform selector composite for a set of related mechanical schools.

CONCLUSIONS

The following conclusions appear warranted from the analysis.

1. Final school grade is a more predictable criterion than Days-in-training.

2. For schools with a final school grade criterion, the present Navy ASVAB composites are about a valid as were Navy Basic Test Battery Composites.

3. A single ASVAB selector composite for both BE/E and "A" schools will reduce BE/E school attrition and still be effective in "A" schools for a group of electromechanical schools.

4. The large number of sets of 2 and 3 tests yielding very similar validities suggests a lack of differential validity among the ASVAB tests.

5. Effective school classification in the Navy may be accomplished with a fewer number of subtests in the ASVAB, through elimination or combining of current ones.

6. Research on school criterion measures with particular emphasis on self-paced schools is needed.

REFERENCES

Christal, R. E. What is the Value of Aptitude Tests? 1975 Miltary Testing Association Proceedings.

Dann, J. E and Abrahams, N. M. Development of a Selector Composite for the Armed Services Vocational Aptitude Battery to Predict Attrition from the Basic Electricity and Electronics School (Unpublished report, 1977), Navy Personnel Research and Development Center, San Diego, CA.

Swanson, L. Validation of the Armed Services Vocational Aptitude Battery (ASVAB form 6 in thirty-one Navy Class A-Schools (unpublished, 1976), Navy Personnel Research and Development Center, San Diego, CA

Thomas, P. J. Basic Test Battery Selection Composite Effectiveness During 1971, Letter Report for Chief of Naval Training, August 1973.

Thomas, P. J. Validation of Form 7 of the USN Basic Test Battery in 108 Class "A" Schools, Naval Personnel and Training Research Laboratory, Technical Bulletin (STB 70-7).

Table 1

Samples Included in ASVAB 6/7 Analysis

Course or Rating		Course Code	Location	N With Criteria Total	Graduate	Academic Drop
Air Controlman	AC	6278	Memphis	52	52	0
Aviation Machinist's Mate, Jet	ADJ	6501	Memphis	385	365	20
Aviation Structural Mechanic, Hydraulics	AMH	6517	Memphis	78	78	0
Aviation Structural Mechanic, Structures	AMS	6518	Memphis	89	89	0
Aviation Ordanceman	AO	6506	Memphis	136	136	0
Avionics Technician, Aviation Electronics Technician	AT	6239	Memphis	265	233	32
Aviation Antisubmarine Warfare Operator	AW	6537	Memphis	92	92	0
Avionics Technician, Aviation Antisub Warfare Technician	AX	6241	Memphis	60	36	4
Aviation Maintenance Administration	AZ	6528	Meridian	66	66	0
Boiler Technician	BT	6260	Great Lakes	753	701	52
Communications Technician, Administrative	CTA	6020	Corry Station	57	48	9
Communications Technician, Communications	CTO	6053	Corry Station	73	60	13
Communications Technician, Collection	CTR	6301	Corry Station	55	39	16
Communications Technician, Technical	CTT	6302	Corry Station	118	89	29
Communications Technician, Field Operations Special Non-Morse	CTT	6320	Corry Station	35	35	0
Dental Technician	DT	6086	San Diego	166	159	7
Electrician's Mate	EM	6070	Great Lakes	169	169	0
Engineman	EN	6261	Great Lakes	389	382	7
Electronics Technician, Communications	ET	6263	Great Lakes	254	254	0
Electronics Technician, Radar	ET	6265	Great Lakes	202	202	0
Electronics Technician, Communications	ET	6266	Great Lakes	64	64	0
Fire Control Technician, Missile	FT	6027	Great Lakes	91	91	0
Gunner's Mate, Guns	GM	6115	Great Lakes	109	109	0
Hospitalman	HM	6084	Great Lakes	1214	1126	88
Hospitalman	HM	6085	San Diego	1079	1021	58
Hull Maintenance Technician	HT	6119	San Francisco	160	158	2
Hull Maintenance Technician	HT	6120	Philadelphia	289	287	2
Machinist's Mate	MM	6262	Great Lakes	1444	1411	33
Mess Management Specialist	MS	6125	San Diego	103	103	0
Operations Specialist	OS	6142	Great Lakes	220	220	0
Postal Clerk	PC	6300	Ft. B. Harrison	36	36	0
Polaris/Poseidon Electronics	PE	6146	Dam Neck	91	91	0
Photographer's Mate	PH	6523	Corry Station	43	43	0

(Continued on next page)

Table 1 (continued)

Samples Included in ASVAB 6/7 Analysis

Course or Rating	Course Code		Location	N With Criteria Total	Graduate	Academic Drop
Personnelman	PN	6102	Meridian	135	124	11
Aircrew Survival Equipmentman	PR	6519	Lakehurst	76	75	1
Quartermaster	QM	6001	Orlando	65	65	0
Radioman	RM	6144	San Diego	681	643	38
Radioman Sea Duty	RM	6380	San Diego	225	225	0
Radioman Shore Duty	RM	6381	San Diego	221	221	0
Signalman	SM	6005	Orlando	42	42	0
Yeoman	YN	6057	Meridian	212	174	38

Table 2

Zero-Order and Multiple Correlation for ASVAB Tests and Composites and the AFQT Against School Performance for 19 Schools Using the General Technical Composite

School	AC		AM		AZ		HM		HM		MS		OS		SM	
School Code	6278		6537		6528		6084		6085		6125		6142		6005	
N	52		92		66		1214		1079		103		220		42	
Criterion	FSC		FSC		FSC		FSC		FSC		FSC		FSC		FSC	
ASVAB Test	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c
GI	17	49	21	25	16	48	26	48	24	53	20	51	24	42	12	41
NO	22	38	29	32	37	64	32	49	31	49	28	47	31	47	-13	10
AD	-20	-32	21	19	20	21	15	23	14	20	03	06	09	16	32	41
WK	27	62	17	25	43	75	46	67	32	64	30	76	16	49	-04	40
AR	25	65	34	39	48	77	41	65	36	66	34	70	39	51	48	65
SP	13	-02	17	20	22	24	07	15	13	27	-15	-17	09	20	10	19
MK	25	51	46	49	60	80	49	66	43	66	42	68	37	56	43	62
MC	44	66	34	38	47	69	30	52	27	55	26	56	45	59	22	48
GS	38	66	39	42	34	65	47	65	41	66	42	70	26	50	07	41
EI	32	56	27	31	47	71	31	52	23	51	39	63	20	35	11	40
SI	50	69	34	37	28	55	15	39	16	46	37	62	26	47	19	40
AI	51	63	23	25	41	53	19	38	18	45	25	47	22	38	23	33
R(12 tests)	67		60		75		60		60		67		62		69	
R(5 tests)	64		58		72		59		58		65		62		64	
Composites																
GT	39	67	23	30	54	80	49	73	44	70	53	72	32	58	31	57
MC	54	73	27	33	52	77	36	61	33	63	50	76	37	60	18	49
EL	41	67	34	38	63	83	51	70	47	71	56	79	25	41	36	59
CL	08	38	31	36	46	71	40	62	37	62	43	67	27	53	11	44
AFQT	29	51	32	35	44	71	35	57	34	63	41	72	31	54	23	53

School	CTO		CTR		CTT		CTT		DT		PN		PC		PN		RM		RM		RM	
School Code	6053		6301		6302		6320		6046		6523		6300		5102		6144		6380		6381	
N	73		55		118		35		166		43		36		135		681		225		221	
Criterion	DAYS		DAYS		DAYS		DAYS		DAYS		DAYS		DAYS		DAYS		DAYS		DAYS		DAYS	
ASVAB Test	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c
GI	-08	-11	11	17	02	-01	05	-06	09	07	23	28	04	01	11	01	10	14	09	21	-01	01
NO	33	27	20	24	02	-01	-03	-17	-07	-06	-10	-06	-07	-08	-18	-23	08	12	07	17	09	09
AD	01	02	34	35	-10	-11	08	-07	-06	-06	06	05	22	19	-13	-17	01	03	01	07	03	04
WK	-01	-08	-48	07	-23	-19	-13	-25	07	05	13	21	13	04	-17	-23	09	14	20	33	-02	01
AR	-05	-10	27	29	15	05	-11	-23	-06	-02	00	12	-17	-12	-02	-13	05	12	11	27	06	06
SP	06	07	25	29	11	10	04	00	-04	-04	03	-07	13	13	04	03	04	04	-03	-06	05	06
MK	00	-07	41	43	07	01	01	-14	06	05	06	14	-02	00	-05	-14	05	11	09	23	04	05
MC	-06	-10	-05	08	03	-04	-05	-16	-07	-05	35	37	20	19	00	-08	04	09	06	20	02	03
GS	-09	-13	06	14	-04	-08	-27	-34	04	04	12	19	05	03	-04	-13	07	12	18	29	04	05
EI	-10	-14	-17	-02	36	04	-23	-29	03	03	28	32	-06	-06	-04	-11	11	15	-02	10	00	07
SI	05	01	12	19	05	01	-06	-12	-06	-04	36	39	25	22	-02	-09	12	15	05	16	00	02
AI	05	-01	26	30	13	12	-22	-15	-14	-10	35	38	-18	-19	09	03	07	11	16	23	06	06
R(12 tests)	50		70		40		50		27		55		73		32		18		31		15	
R(5 tests)	45		65		37		45		25		51		64		31		17		30		14	
Composites																						
GT	-24	-09	10	18	-03	-04	-15	-26	01	02	10	19	-02	-04	-11	-18	08	14	18	31	02	04
MC	-01	-08	01	12	-06	-08	-11	-24	-04	-01	38	36	30	19	-08	-15	10	15	12	28	00	02
EL	-08	-12	20	24	09	01	-02	-18	03	02	18	24	-08	-08	-05	-15	08	16	11	27	04	05
CL	19	08	26	29	-13	-13	-09	-22	-03	-02	07	15	14	07	-21	-26	08	13	12	26	05	06
AFQT	01	-05	04	13	05	00	-06	-20	02	02	08	17	12	06	-09	-16	01	08	14	27	03	05

Notes:
1. Decimal points are omitted.
2. Validities of the composite used for selection and the tests in it are underlined.
3. A negative validity is expected for the DAYS criterion.
4. r_u = Uncorrected validities. r_c = Validities corrected for restriction of range.

Table 2 (Continued)

Zero Order and Multiple Correlations for ASVAB 6/7 and Composites and the AFQT Against School Performance

Mechanical Composite

School	EH		GM		BT		EN		HT		HT		MM		PR	
School Code	6070		6155		6260		6261		6119		6120		6262		6519	
N	169		109		753		389		160		289		1444		76	
Criterion	FSC		FSG		DAYS		DAYS		DAYS		DAYS		DAYS		DAYS	
Composite Used	-		-		-		-		-		-		-		-	
ASVAB Test	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c
GI	51	67	32	59	-11	-22	-20	-34	-17	-36	-07	-17	-29	-45	-25	-13
NO	29	38	13	33	-17	-24	-10	-21	-21	-30	-10	-16	-27	-39	-16	-16
AD	08	08	05	26	-04	-06	-13	-17	-14	-24	06	03	-13	-18	-01	-01
WK	46	73	41	70	-08	-25	-14	-33	-07	-31	-08	-19	-31	-49	27	19
AR	56	76	25	60	-19	-31	-20	-35	-18	-35	-15	-23	-39	-54	-09	-09
SP	29	53	23	52	-01	00	-02	-03	-15	-27	06	08	-15	-28	-11	-11
MK	57	74	12	44	-19	-30	[illegible]	-33	-20	-35	-13	-21	-35	-50	-18	-16
MC	58	79	32	66	-12	-27	[illegible]	-37	-10	-32	-14	-24	-26	-47	-10	-08
GS	55	81	42	69	-11	-26	-13	-31	-22	-39	-09	-20	-30	-48	-15	-14
EI	43	68	36	64	-15	-27	-10	-36	-15	-33	-18	-26	-27	-46	-05	-05
SI	46	74	28	64	-09	-26	-09	-30	-21	-39	-03	-17	-23	-44	-17	-13
AI	30	55	46	69	-18	-29	-17	-33	-16	-33	-11	-20	-15	-34	-15	-13
R(12 tests)	78		58		29		33		35		29		44		57	
R(5 tests)	76		57		29		32		32		27		44		55	
Composites																
GT	60	80	40	71	-17	-31	-21	-37	-16	-36	-14	-24	-40	-56	07	03
MC	67	84	46	73	-14	-29	-21	-37	-19	-39	-11	-21	-35	-53	-02	-04
EL	72	86	41	71	-23	-34	-25	-40	-25	-41	-19	-27	-40	-55	-17	-13
CL	37	57	23	57	-14	-26	-18	-33	-20	-37	-04	-15	-32	-48	02	00
Special	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AFQT	59	79	38	69	-16	-30	-15	-32	-12	-34	-05	-16	-36	-52	03	00

Electronics Composite

School	AO		ET		ET		ET		FT		ADJ		AT		AX	
School Code	6506		6263		6265		6266		6027		6501		6239		6241	
N	136		254		202		64		91		385		265		60	
Criterion	FSG		FSG		FSG		FSG		FSG		DAYS		DAYS		DAYS	
Composite Used	-		-		-		-		-		-		-		-	
ASVAB Test	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c
GI	27	66	10	48	13	53	05	16	21	54	-22	-46	-05	-31	-24	-53
NO	15	20	11	35	13	47	11	18	02	28	-11	-33	-15	-23	-01	-20
AD	-05	-10	02	07	13	26	-05	-04	02	13	-14	-18	-02	05	-07	-05
WK	28	69	24	58	19	65	11	23	32	65	-26	-50	-11	-40	-30	-57
AR	37	76	28	66	27	70	06	21	19	69	-25	-63	-22	-48	-24	-53
SP	25	16	30	56	27	58	17	26	28	58	15	20	-12	-28	-02	-34
MK	15	67	32	68	33	77	15	27	17	65	-29	-55	-20	-47	-15	-49
MC	45	77	36	67	45	76	16	27	30	72	-16	-44	-21	-47	-13	-47
GS	22	71	28	67	36	76	19	29	39	77	-24	-52	-21	-49	-05	-44
EI	20	67	36	69	38	77	-03	15	35	75	-14	-44	-10	-43	-22	-53
SI	30	70	23	57	28	63	12	23	24	62	-24	-49	-08	00	-21	-52
AI	28	61	25	55	24	56	09	20	30	63	-25	-48	-08	-35	-18	-49
R(12 tests)	57		50		56		41		55		45		37		51	
R(5 tests)	55		49		55		37		52		43		35		49	
Composites																
GT	41	78	33	68	29	73	11	24	35	75	-32	-57	-20	-48	-42	-62
MC	48	80	39	71	42	79	17	28	39	76	-29	-55	-17	-47	-30	-57
EL	41	79	46	75	52	82	14	26	47	80	-36	-60	-26	-52	-25	-55
CL	15	47	14	43	21	60	04	14	16	49	-26	-48	-14	-32	-20	-45
Special	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AFQT	42	76	35	68	38	74	20	29	37	74	-20	-48	-21	-49	-21	-30

Clerical Composite / Special Composites

School	CTA		YN		AMH		AMS		QM	
School Code	6020		6057		6517		6518		6001	
N	57		212		78		89		65	
Criterion	DAYS		DAYS		FSG		FSG		FSG	
Composite Used	-		-		WK+MC		WK+MC		AR+SI	
ASVAB Test	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c	r_u	r_c
GI	26	28	09	07	41	61	27	54	58	73
NO	-08	-14	-13	-15	04	20	31	42	33	45
AD	02	-07	-07	-10	07	25	-16	-11	27	32
WK	-08	-13	04	01	31	61	22	58	66	78
AR	07	01	10	07	38	62	49	70	76	88
SP	-16	-18	-07	-08	06	-16	-03	-03	10	21
MK	07	00	06	03	29	51	46	65	60	73
MC	-08	-07	04	02	47	70	36	66	46	70
GS	07	06	09	07	29	53	31	61	54	72
EI	11	13	02	00	47	59	24	54	51	71
SI	08	07	04	03	43	59	44	67	38	70
AI	28	30	08	07	44	60	38	61	43	66
R(12 tests)	48		23		69		67		84	
R(5 tests)	45		22		67		65		83	
Composites										
GT	02	-05	09	04	43	68	45	70	79	89
MC	-02	-05	07	03	63	79	47	71	68	84
EL	10	06	11	06	51	70	54	74	71	85
CL	-07	-13	-07	-11	18	50	16	45	59	72
Special	-	-	-	-	53	74	40	68	67	84
AFQT	05	-02	06	03	45	66	36	64	66	82

NOTES:
1. Decimal Points are omitted.
2. Validities of the composite used for selection and the tests in it are underlined.
3. A negative validity is expected for the DAYS criterion.
4. r_u = uncorrected validities. r_c = Validities corrected for restriction of range.

Table 3

Validities of ASVAB 6/7 Selection Composites and the Most Valid Sets of Two and Three ASVAB Tests for 41 Schools

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (WK+AR) r_u	r_c	Most Valid Sets of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
AC	6278	52	FSG	39	67	AR+AI	54	73	MC+SI+AI	55	73
						SI+AI	54	72	WK+MC+SI	54	73
						MK+AI	51	70	AR+MC+AI	54	73
						AR+SI	50	71	AR+GS+AI	54	73
AW	6537	92	FSG	23	30	NO+MK	41	44	MK+EI+GS	36	40
						MK+GS	37	40	NO+AD+MK	35	39
						MK+EI	36	40	AR+MK+GS	34	38
						NO+WK	33	38	MK+EI+MC	33	38
									MK+MC+SI	33	38
AZ	6528	66	FSG	54	80	MK+EI	68	85	MK+EI+MC	65	83
						WK+MK	66	84	MK+EI+GS	62	82
						MK+MC	62	81	MK+MC+AI	61	81
						AR+MK	61	82	AR+MK+GS	60	82
						NO+MK	61	81			
HM	6084	1214	FSG	49	73	WK+MK	55	72	WK+AR+GS	52	71
						MK+GS	53	71	AR+MK+GS	52	71
						NO+WK	50	69	MK+EI+GS	50	70
HM	6085	1079	FSG	44	70	WK+MK	49	72	AR+MK+GS	49	72
						MK+GS	47	70	WK+AR+GS	48	71
						NO+WK	45	69	MK+EI+GS	44	69
						NO+WK	45	67	WK+AR+MC	43	69
						AR+MK	44	69			
MS	6125	103	FSG	53	79	NO+WK	58	79	MK+EI+GS	56	78
						WK+MK	55	79	WK+AR+GS	55	80
						MK+EI	53	76	WK+AR+SI	52	78
						MK+GS	51	76	AR+EI+GS	52	78
									AR+MK+GS	52	78
OS	6142	220	FSG	32	58	MK+MC	45	64	MK+MC+SI	43	63
						AR+MC	44	64	WK+AR+MC	41	62
						NO+AR	41	62	AR+EI+MC	41	62
						AR+MK	40	62	AR+MC+SI	41	62
						NO+MK	39	59	MK+EI+MC	41	61
									MK+MC+AI	41	61
SH	6005	42	FSG	31	57	AR+MK	51	67	AD+WK+AR	44	63
						AR+SI	47	65	AR+MC+SI	42	62
						AD+AR	47	64	AR+MK+GS	41	62
						AD+MK	47	63	MK+MC+SI	37	59
						AR+MC	43	63			

Table 3 (Continued)

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (WK+AR) r_u	r_c	Most Valid Sets Of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
CTO	6053	73	DAYS	-04	-09	MK+EI	-06	-11	AR+EI+MC	-09	-13
						AR+MC	-06	-11	MK+EI+GS	-08	-12
						SP+EI	-05	-09	MK+EI+MC	-07	-11
						MK+GS	-05	-11			
CTR	8301	55	DAYS	10	18	None valid.			None valid.		
CTT	6302	118	DAYS	-03	-06	AD+WK	-20	-18	WK+AD+NO	-13	-13
						NO+WK	-13	-13			
						WK+MC	-12	-12			
CTT	6320	35	DAYS	-15	-26	AR+AI	-23	-23	AR+GS+AI	-30	-36
						SI+AI	-18	-19	AR+EI+GS	-29	-35
						MK+GS	-14	-26	WK+AR+GS	-21	-30
						MK+EI	-14	-24	MK+EI+GS	-21	-30
						MK+AI	-14	-23			
DT	6086	166	DAYS	01	02	AR+AI	-13	-07	NO+AD+SP	-11	-11
						AD+SP	-11	-11	AR+SP+MC	-11	-07
						SP+MC	-11	-09	AR+MC+AI	-11	-07
						AR+SP	-11	-08	MC+SI+AI	-11	-07
						SI+AI	-11	-08			
PH	6523	43	DAYS	10	19	None valid.			None valid.		
PC	6300	36	DAYS	-02	-04	AR+AI	-22	-20	AR+GS+AI	-18	-15
						MK+AI	-16	-16	AR+MC+AI	-10	-10
						NO+AR	-13	-12	AR+EI+GS	-09	-09
						GI+AI	-12	-12			
						AR+MK	-10	-09			
PN	6102	135	DAYS	-11	-18	NO+WK	-22	-26	WK+AD+NO	-21	-26
						AD+WK	-19	-24	AD+WK+AR	-16	-22
						NO+AD	-17	-22	NO+AD+MK	-16	-21
									NO+AD+AR	-15	-21
RM	6144	681	DAYS	08	14	None valid.			None valid.		
RM	6380	225	DAYS	18	31	None valid.			None valid.		
RM	6381	221	DAYS	02	04	None valid.			None valid		

Table 3 (Continued)

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (WK+MC+SI) r_u	r_c	Most Valid Sets of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
EM	6070	169	FSG	67	84	MK+MC	69	85	MK+MC+SI	72	87
						MK+GS	69	84	AR+MK+GS	71	85
						AR+MC	67	84	AR+EI+GS	70	85
						WK+MC	65	84	AR+GS+AI	70	85
									MK+EI+GS	70	85
									MK+EI+MC	69	85
									2MK+AR+GS	68	83
GM	6115	109	FSG	46	73	AR+AI	49	74	AR+GS+AI	51	75
						SI+AI	48	73	AR+MC+AI	49	75
						GI+AI	47	73	MC+SI+AI	49	74
						WK+MC	45	73	WK+AR+GS	48	74
									WK+AR+GS	48	74
									2MK+AR+GS	29	62
BT	6260	253	DAYS	-14	-29	MK+AI	-25	-35	MK+MC+AI	-21	-33
						AR+AI	-24	-35	AR+MK+GS	-21	-33
						NO+AR	-22	-34	AR+MC+AI	-20	-33
						NO+MK	-22	-33	AR+GS+AI	-20	-32
						MK+EI	-22	-33	MK+EI+MC	-20	-32
						AR+MK	-21	-33	2MK+AR+GS	-21	-33
						MK+MC	-20	-32			
EN	6261	389	DAYS	-21	-37	AR+AI	-25	-40	AR+EI+MC	-24	-39
						GI+AI	-25	-40	AR+MC+AI	-24	-39
						AR+MC	-25	-40	AR+MK+GS	-23	-38
						MK+AI	-24	-39	MK+MC+AI	-23	-38
						MK+MC	-23	-39	AD+WK+AR	-23	-38
						MK+EI	-23	-38	2MK+AR+GS	-22	-38
						AR+MK	-23	-38			
HT	6119	160	DAYS	-19	-39	MK+GS	-25	-41	AR+GS+AI	-26	-42
						MK+AI	-24	-40	MK+EI+GS	-25	-41
						AR+SI	-24	-41	AR+MK+GS	-25	-41
						AR+AI	-23	-40	AR+EI+GS	-24	-41
						NO+MK	-23	-37	AR+SP+GS	-23	-40
									MK+MC+AI	-23	-40
									NO+AD+AR	-23	-37
									NO+AD+MK	-23	-37
									2MK+AR+GS	-24	-40
HT	6120	289	DAYS	-11	-21	AR+MC	-17	-26	AR+EI+MC	-21	-28
						AR+AI	-17	-25	MK+EI+MC	-21	-28
						MK+AI	-17	-25	AR+EI+GS	-19	-27
						MK+MC	-16	-25	AR+MC+AI	-18	-26
									MK+EI+GS	-18	-26
									MK+MC+AI	-18	-26
									2MK+AR+GS	-16	-24

Table 3 (Continued)

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (WK+MC+SI) r_u	r_c	Most Valid Sets of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
MM	6262	1444	DAYS	-35	-53	AR+MC	-39	-55	WK+AR+MC	-41	-56
						AR+MK	-39	-54	WK+AR+SI	-40	-56
						NO+AR	-39	-53	WK+AR+GS	-40	-55
						WK+MK	-38	-54	AR+MK+GS	-40	-55
						MK+EI	-37	-53	AR+EI+MC	-39	-55
						AR+SI	-37	-54	AR+EI+GS	-39	-55
						NO+WK	-37	-52	2MK+AR+GS	-39	-54
						MK+MC	-36	-53	AR+MC+SI	-38	-54
						MK+GS	-36	-53	AR+GS+AI	-38	-54
						NO+MK	-36	-51			
						AR+AI	-36	-53			
						MK+AI	-34	-51			
PR	6519	76	DAYS	-02	-04	MK+GS	-21	-17	MK+MC+SI	-19	-19
						MK+AI	-19	-16	AR+MK+GS	-18	-15
						NO+WK	-19	-17	AR+GS+AI	-18	-14
						SI+AI	-18	-14	MK+EI+GS	-18	-14
						GI+AI	-18	-15	MK+MC+AI	-18	-13
						NO+SP	-17	-17	SP+MK+MC	-17	-13
						SP+MK	-17	-15			

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (AR+MK+EI+GS) r_u	r_c	Most Valid Sets Of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
AO	6506	136	FSG	41	79	AR+MC	49	81	WK+AR+MC	50	81
						WK+MC	47	80	AR+EI+MC	48	81
						AR+AI	42	79	WK+MC+SI	48	80
						MK+MC	41	79	AR+MC+AI	46	80
									AR+MC+AI	46	80
									MK+MC+SI	46	80
									2MK+AR+GS	32	77
ET	6263	254	FSG	46	75	MK+EI	44	74	MK+EI+MC	47	75
						MK+MC	42	73	AR+EI+MC	45	74
						WK+MC	41	72	AR+EI+GS	44	74
						AR+MC	39	71	MK+EI+GS	44	74
									SP+MK+EI	43	73
									2MK+AR+GS	40	73
ET	6265	202	FSG	52	82	MK+EI	51	82	MK+EI+MC	54	83
						MK+MC	50	81	MK+EI+GS	52	82
									AR+EI+MC	52	82
									AR+EI+GS	51	82
									2MK+AR+GS	43	80

Table 3 (Continued)

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (AR+MK+EI+GS)		Most Valid Sets Of Two ASVAB Tests			Most Valid Sets of Three ASVAB Tests		
				r_u	r_c		r_u	r_c		r_u	r_c
ET	6266	64	FSG	14	26	SP+MK	21	30	AR+SP+GS	22	31
						MK+GS	21	30	SP+MK+MC	22	31
						SP+MC	20	29	WK+SP+MK	21	31
						MK+MC	20	29	MK+MC+SI	19	29
						WK+SP	19	28	2MK+AR+GS	18	28
						NO+SP	19	27			
FT	6027	91	FSG	47	80	MK+EI	43	79	MK+EI+GS	48	81
						MK+GS	40	79	AR+EI+GS	45	80
									AR+GS+AI	45	80
									AR+SP+GS	45	80
									WK+AR+GS	44	80
									2MK+AR+GS	36	77
PE	6146	99	FSG	39	76	MK+GS	46	78	MK+EI+GS	45	77
						MK+EI	45	77	MK+EI+MC	44	77
						MK+MC	43	76	2MK+AR+GS	40	76
						SP+MK	33	71	SP+MK+EI	39	75
						WK+MK	32	72	MK+MC+SI	37	74
									AR+MK+GS	36	75
ADJ	6501	365	DAYS	-36	-60	WK+MK	-36	-59	AR+MK+GS	-36	-60
						MK+MC	-34	-58	2MK+AR+GS	-36	-59
						MK+GS	-33	-58	AR+GS+AI	-33	-58
									AD+WK+AR	-33	-56
AT	6239	265	DAYS	-26	-52	MK+MC	-26	-51	AR+MK+GS	-29	-53
						MK+GS	-26	-52	2MK+AR+GS	-28	-53
						AR+MK	-25	-51	AR+SP+GS	-25	-51
						AR+MC	-25	-51	WK+AR+MC	-24	-51
						NO+AR	-24	-46	AR+EI+MC	-24	-51
									AR+EI+GS	-24	-51
AX	6241	60	DAYS	-25	-55	WK+AR	-42	-62	AD+WK+AR	-31	-56
						WK+MK	-33	-58	WK+MC+SI	-30	-57
						WK+MC	-29	-56	WK+AR+SI	-24	-55
						WK+SP	-27	-56	2MK+AR+GS	-20	-53

Table 3 (Continued)

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (WK+AD+NO) r_u	r_c	Most Valid Sets Of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
CTA	6020	57	DAYS	-07	-13	NO+SP	-13	-17	WK+SP+MK	-08	-13
						WK+SP	-13	-16	SP+MK+MC	-08	-11
						SP+MC	-12	-14			
						NO+WK	-11	-15			
YN	6057	212	DAYS	-07	-11	NO+SP	-13	-16	NO+AD+SP	-12	-15
						AD+SP	-09	-12			

School	Code	N	Criterion	Validitiy of ASVAB 6/7 Selector Composite (WK+MC) r_u	r_c	Most Valid Sets Of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
AMH	6517	78	FSC	53	74	AR+MC	51	72	WK+MC+SI	63	79
						AR+AI	51	70	AR+EI+MC	59	76
						MK+AI	49	67	AR+MC+SI	57	75
						MK+EI	48	65	AR+MC+AI	57	75
						SI+AI	48	64	MK+EI+MC	57	75
						MK+MC	47	69	MK+MC+SI	57	75
									MK+MC+AI	56	75
AMS	6518	89	FSG	40	68	MK+AI	58	76	AR+MK+GS	56	75
						AR+MK	55	74	2MK+AR+GS	55	74
						AR+AI	54	74	MK+MC+SI	52	73
						MK+MC	53	73	MK+MC+AI	50	72
									AR+MC+SI	49	72
									AR+MC+AI	49	72
									AR+GS+AI	49	72
									WK+MC+SI	47	71

School	Code	N	Criterion	Validity of ASVAB 6/7 Selector Composite (AR+SI) r_u	r_c	Most Valid Sets Of Two ASVAB Tests	r_u	r_c	Most Valid Sets of Three ASVAB Tests	r_u	r_c
QM	6001	65	FSG	67	84	WK+AR	79	89	WK+AR+MC	76	88
						AR+MK	72	85	WK+AR+SI	76	88
						WK+MK	69	82	WK+AR+GI	73	86
						WK+MC	68	83	AD+WK+AR	73	85
						AR+MC	67	84	AR+EI+GS	70	85
						AR+AI	67	85	WK+MC+SI	68	84
						NO+AR	65	80	2MK+AR+GS	68	83
									AR+MK+GS	71	85
									AR+EI+MC	68	84

Notes: 1. Decimal points are omitted from correlations.
2. A negative validity is expected for the DAYS criterion.

Table 4

Validities of Operational and Selected Alternate Composites
in 4 Ratings Presently Using Electromechanical Selectors

Composite	Sample Source		CM (6115)		EM (6[illegible]0)		IC (6076)		CE (6289)	
			r	N	r	N	r	N	r	N
WK+MC+SI (Operational Composite)	BE&E School	(a)	.17	238	.11	230	.34	87	.65	21
		(b)	.28	297	.28	297	.28	297	.28	297
	A-School		.46	109	.67	169				
2MK+AR+GS (New BE&E Composite)	BE&E School	(a)	.39	119	.41	115	.54	44	.79	10
		(b)	.48	297	.48	297	.48	297	.48	297
	A-School		.29	109	.68	169				
AR+MK+CS+EI (Current EL Composite)	BE&E School	(a)	.37	238	.25	230	.51	87	.85	21
		(b)	.46	297	.46	297	.46	297	.46	297
	A-School		.41	109	.72	169				

Note.

1. Validities in row (a) are based on BE&E students who were subsequently assigned to each specific A-School. Validities in row (b) are based on all BE&E students for whom the WK+MC+SI selector composite was used.

2. The validities enclosed in dotted lines are for the composites recommended for each rating.

Table 5

A-Schools Other Than Those Originating With BE/E School For Which Selector Composite Changes Are Proposed

Rating		Code	Criterion	Present Selector	Validity r_u	r_c	Proposed Selector	Validity r_u	r_c
Personnelman	PN	6102	DAYS	WK+AR	-11	-18	MO+WK	-22	-26
Boiler Technician	BT	6260	DAYS	WK+MC+SI	-14	-19	MK+AI	-25	-35
Engineman	EN	6261	DAYS	WK+MC+SI	-21	-37	MK+AI	-24	-39
Machinists Mate	MM	6262	DAYS	WK+MC+SI	-35	-53	MK+AI	-34	-53
Aircrew Survival Equipmentman	PR	6519	DAYS	WK+MC+SI	-02	-04	MK+AI	-19	-16
Aviation Structural Mechanic, Hydralics	AMH	6517	FSG	WK+MC	53	74	AR+MC+AI	57	75
Aviation Structural Mechanic, Structures	AMS	6518	FSG	WK+MC	40	68	AR+MC+AI	49	72
Quartermaster	QM	6001	FSG	AR+SI	67	84	WK+AR	79	89

ASSESSMENT OF ARMED SERVICES VOCATIONAL APTITUDE BATTERY VALIDITY

Lonnie D. Valentine
Air Force Human Resources Laboratory
Personnel Research Division
Brooks Air Force Base, Texas

It is my first intention today to provide you a capsule description of the Air Force generated validity information available on the Armed Services Vocational Aptitude Battery (ASVAB) along with appropriate references to such studies for those of you who wish or need more detailed data from the studies. I shall also briefly discuss the need for validations against job performance and suggest one possible approach. Last, I shall discuss the kinds of joint service test validation studies which are needed and describe some of the data requirements and pitfalls associated with them.

ASVAB is required to serve a broad range of purposes which are not always compatible in a relatively short battery. For the services, it must provide a measure for enlistment qualification (i.e., a selection measure) and in addition must provide the initial classification measures required by the various services. With respect to classification, the philosophies and requirements of the services differ. Moreover, the battery is used in a High School testing program which covers the range of grades 9 through 12. For this purpose, maximization of differential aptitude information is of prime importance. By contrast, the service programs require balance between maximum prediction and maximum differentiation between aptitude areas. I'll return to this later.

In 1973, Vitola, Mullins, and Cross reported on validity for prediction of Air Force technical training grades of Air Force classification composites derived from ASVAB, Form 1, and compared validity of these composites with that of their counterparts from the Airman Qualifying Examination (AQE) (see AFHRL-TR-73-7). Their subjects had been enlisted on the basis of AQE and were administered ASVAB-1 at Lackland Air Force Base on their sixth day of Basic Military Training.

They found that, in general, the obtained validity of ASVAB composites was equal to or slightly higher than that of their AQE counterparts; the small discrepancy in favor of the ASVAB composites probably was an artifact of assignment to training courses (and consequently direct restriction of range) on the AQE composites. They also found that, for most courses, the relevant ASVAB composite was more valid for that course than were the other three composites. Validities from this study are summarized in Tables 1 through 4 of the handout.

Valentine (1977) reported a large scale validation of ASVAB-3 Air Force composites against Air Force technical training grades. The objectives of the study were to (a) investigate validity of the Armed Services Vocational Aptitude Battery and of educational background data for Air Force technical training, (b) investigate unique predictive contribution of both educational background and test data for Air Force technical training success, and (c) assess homogeneity of prediction equations for subgroups defined by sex and race.

ASVAB-3 data were collected on Air Force non-prior service enlisted accessions for September 1973 through October 1975. Analyses were conducted on 43 clusters of enlisted technical courses; formation of these clusters was based on a count of available cases in various courses and clustering together of those which were judged to be similar. The major criterion variable was final course grade.

Variables used in the study were (a) an Armed Forces Qualification Test (AFQT) score and four Air Force Aptitude Indexes (Mechanical, Administrative, General, and Electronics) all derived from ASVAB-3, (b) a series of 41 binary variables indicating successful completion or non-completion of specific high school courses, (c) disposition from training (graduation or elimination), (d) final course grade, (e) ethnic identity (Caucasian, Black, or Other Minority), (f) sex (male or female), and (g) course cluster identity.

Half of the male Caucasians in each of the 43 course clusters were randomly selected as an Educational Index (EI) development sample. The EI development sample was restricted to Caucasian males because, in most clusters, inclusion of women and ethnic minorities would have reduced their number in the remaining sample below a desirable number for the analyses contemplated for it. For each of the 43 course clusters, an Education Index was developed using the EI development sample. For this purpose, the sample was divided into an upper and a lower 50% criterion dichotomy by assigning all fail cases to the lower group along with enough of the cases with the lowest final course grades to complete 50% of the development sample. The 41 course completion variables were item analyzed against this course success dichotomy. Those courses with significant positive correlation with the criterion were assigned an EI scoring weight of + 1 while those with significant negative correlation with the criterion were assigned a scoring weight of - 1. The EI development samples were excluded from all subsequent analyses; thus, all validities reported in the study represent cross-validation values.

Table 5 of the handout summarizes the areas for which analyses were accomplished and shows the number of cases available in the cross-validation samples. For these samples, Table 6 shows validity of the AFQT, the four Air Force aptitude composites, and the Education Index against final course grade. Table 7 summarizes tests of hypotheses about independent contribution

to prediction of test and of educational background information. Both test and educational background data demonstrated usefulness for prediction of technical training performance; moreover, when used in combination with each other, more accurate predictions are achieved than through the use of either alone. Generally, of the two kinds of data, test data alone provided more accurate predictions than did educational data alone, and, moreover, introduction of test data to an equation based on educational background provided a larger increase in prediction accuracy than was achieved with introduction of educational background into a test-based prediction equation. These observations also hold for prediction equations based on specific race or sex subsamples.

To test hypotheses about homogeneity of separate race or sex regression equations, a series of regression problems involving race membership, sex membership, AFQT, the Selector AI, the Educational Index, and interactions of race or sex with the test and educational variables as predictors of final course grade were computed and compared via the $\underline{F}$ statistic. Tables 8 and 9 summarize these hypotheses and the tests of them. While not tabled in your handout, these same hypotheses were also tested for the Educational Index and for the test variables separately.

In many instances, separate race or sex prediction equations are not homogeneous (i.e., the subgroup equations differ from each other enough that added accuracy in prediction is achieved by using a separate equation for each subgroup); this observation is more often true for race based subgroups and for predictions based on educational background data. In all but two instances, there were significant differences in the separate race equations for predicting technical training performance from educational background. In most instances, the data suggest that differences in race-based prediction equations are attributable to the equations' intercepts; that is, while usually the predicted technical training grade increases for each subgroup by about the same amount for each increase of one score unit on the predictor, the constants added into the equations differ. This results in parallel prediction lines for the subgroups which differ mainly in level.

Table 10 of your handout demonstrates the impact of these equation differences. This table was developed from separate subgroup (i.e., Causasian, Black, male, female) regression equations for predicting training performance from test and educational background data. From this table it can be seen that, when total group means on the selector AI, AFQT, and EI are substituted into the Black and Caucasian equations, a lower criterion value is predicted by the Black equation. Thus, when a single overall equation is used, the tendency would be to predict higher Black criterion performance than is observed. A single overall equation tends to underpredict female criterion performance in Food Service, Administrative, and Medical specialties and to overpredict for them in Mechanical specialities and in Law Enforcement.

One validation study has been accomplished specifically for the benefit of the high school testing program. This provided validity information against high school vocational-technical curriculum grades (Jensen & Valentine, 1976). The sample for this study consisted of approximately 4,300 high school students primarily from the northeastern sector of the country. Each of these students was enrolled in one or another of 41 different high school vocational-technical courses. Each student had been administered ASVAB-2 during the 1973-74 school year. High schools participating in the study provided the final course grade for each student for the vocational-technical course in which the student was enrolled. For each of the 41 high school vocational-technical courses, validities of the nine subtests of ASVAB-2 against the course grade were computed and multiple correlations against that same criterion were obtained. Table 11 presents the average validities of the ASVAB subtests for vocational-technical courses judged to be subsumed under the four aptitude areas used in the Air Force classification program. For these 41 vocational-technical courses, multiple correlations of the nine ASVAB subtests against course grade ranged from .30 to .93 with a median of .54.

During the spring of 1977, new ASVAB composites for the High School Testing program were developed for school year 1977-78. These composites were based on an oblique factor analysis of the subtests of the battery and are believed to be more representative of dimensions of human ability than were the composites which they replace. A primary goal in developing these new composites was to provide a set of scores for high school counseling which provide better differentation among abilities than was available from the previous composite set reported to the schools by MEPCOM.

For an unpublished study (Valentine and Mathews), a USAREC computer tape file of data on AFEES testing accomplished 1 April 1976 thru 31 March 1977 was obtained from the Defence Manpower Analysis Center. This file identifies individuals processed at the AFEES, indicates the service for which they were processed, identifies the specific test or tests (by form) administered to them, and, in the case of ASVAB, records all subtest raw scores. A subfile, developed from the larger file, consisted of individuals AFEES processed for the Air Force who were administered Form 5, 6, or 7 of ASVAB. The file was further reduced by deletion of cases on whom identification data was insufficient for collation with other files or for whom test score data in the file was incomplete; effectively, this was a file consisting of Air Force personnel who were administered Form 6 or 7 of the ASVAB.

The resulting file was matched against the Air Force Technical Training file to obtain identity of technical training courses and final course grades. This collated file was subdivided on the basis of ASVAB form taken (6 or 7) and technical training courses completed. Analyses were accomplished separately on resulting samples when N was equal to or greater than 50. For Form 6, 10 such samples were available and for Form 7, 16 samples were available.

For each case, raw scores for the High School composites in use during the 76-77 school year and for the High School composites proposed for school year 77-78 were computed. In each of the 26 available samples, correlations of these two sets of composite scores with final course grade were computed.

Table 12 lists the 16 technical courses for which validation samples were available; the first column of Table 1 shows the AFSC associated with the course, column 2 shows the associated job title, while the last column indicates the Air Force composite and percentile normally required for assignment to the course.

Table 13 summarizes validity and composite intercorrelational information for both the current and proposed High School composites for ASVAB Form 6 samples, while Table 14 provides a similar summary for Form 7 samples. For each sample, the tables show the number of cases in the sample, the range of composites vs. final course grade correlations obtained for each set of composites separately, and the median intercorrelation within each set of composites. It should be noted that correlational values presented are obtained values; they have not been corrected for range restriction.

The most immediate conclusions from this data are that proposed High School ASVAB composites (a) should prove as useful as the current set for success predictions and (b) should prove more useful than the present set for counseling use, both in terms of greater spread of validities among the composites (therefore providing easier identification of one or two relevant composites) and in terms of greater differentiation in ability patterns for individual subjects (as reflected in lower intercorrelations among the composites in the set).

I mentioned earlier that ASVAB is expected to serve a variety of uses which necessitate the battery's use with examinees ranging from ninth graders through seniors and young adults applying for service enlistment. A great deal of evidence suggests that there is a large difference in difficulty of test material between the tenth and eleventh grades. Thus, tests which are ''easy'' enough for ninth and tenth graders are too easy for the other groups, and tests which are of appropriate difficulty for eleventh and twelfth graders are far too hard for ninth and tenth graders. Moreover, the various services have different cut-off requirements. In a 20-item scale, it is near impossible to balance these varied difficulty requirements. This, in turn, impacts on the amount of chance (or unreliability) variance for less mature subjects and unduly restricts variance among the more mature subjects. This, in turn, will tend to limit the battery's validity. One solution to this problem would be production of the battery in ''easy'' forms appropriate for ninth and tenth graders and some of the less demanding service needs, and ''hard'' forms appropriate for eleventh and twelfth grade use and more demanding service selection and classification needs. Judicious normative procedures could calibrate the two versions to service normative standards with a region of normative overlap. This would effectively double battery length and provide more reliable test data

provided procedures (such as a short version placement scale) are employed for service applicants.

Several years ago, I tried out an idea for test validation against operational criteria which I believe may have utility in evaluating the battery against ''job performance.'' I did not formally report the effort or attempt to replicate it on later samples because changes in the standards and criteria by which Air Force career progress occurs, instituted at about that time, confound the criterion. However, there may be places within the other services where the approach is feasible. At that time, Air Force promotion opportunity was limited by such factors as time in service, specialty, and skill level. Working from the Uniform Airman Record file, I sorted several of the more populous specialties into homogeneous subsamples with respect to these factors. I then tested for significant differences in mean selector aptitude index for higher and lower ranking personnel within these year and skill level groups. Table 15 summarizes outcomes for Security Police; similar results were obtained for the several other specialties examined. One can assume that supervisors tend to work for promotion of their most capable workers first. These data certainly provided evidence that, other things being equal, the higher aptitude personnel are the first promoted. For the benefit of any of you who may feel uneasy about this ''backwards'' sort of application of the $\underline{F}$ ratio, let me point out that, in this instance, $\underline{F} = t^2$. Certainly in our future joint ASVAB validation efforts, approaches of this sort might be considered whenever we locate samples whose promotion is not contingent on application of a promotion score equation.

There is continuing concern with DoD for development of a single set of ASVAB composite scores with applicability across programs. Analyses, designed to establish feasibility of a single composite set, are being designed at the present time with ASVAB Forms 6 and 7. Essentially, these will involve cross-service application of current sets and examination of alternatives to these. These later analyses, involving establishment of regression equations for various service schools from subtest information and assessment of cross-service homogeneity, will entail a number of special problems which will require careful control. Among these, extent to which service training criteria have the same meaning, even for essentially identical jobs, is unknown. Thus, we anticipate that these analyses will be relatively complex.

In summary, most Air Force studies indicate useful ASVAB validity for training criteria. Next, validation efforts will explain extent to which a single set of composites, adequately responsive to selection and classification strategies of all the services, are possible.

REFERENCES

Jensen, H.E., & Valentine, L.D., Jr. Validation of ASVAB-2 against civilian vocational-technical high school criteria (AFHRL-TR-76-16). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, March 1976.

Valentine, L.D., Jr. Prediction of Air Force technical training success from ASVAB and educational background (AFHRL-TR-77-18). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, May 1977.

Valentine, L.D., Jr., & Mathews, J.J. Validation of high school composites from Armed Services Vocational Aptitude Battery (ASVAB) (Forms 6 and 7) for Air Force technical training. Unpublished staff report, 23 May 1977, Air Force Human Resources Laboratory, Brooks AFB TX.

Vitola, B.M., Mullins, C.J., & Croll, P.R. Validity of Armed Services Vocational Aptitude Battery, Form 1, to predict technical school success (AFHRL-TR-73-7). Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, July 1973.

Table 1. **ASVAB Validities for Techinical School Courses in the General Cluster**

Course		Minimum AI	N	Mechanical r	Administrative r	General r	Electronics r
				Aptitude Index Validities			
3ABR20630	Imagery Interpreter Specialist	80	116	.37	.30	.42	.43
3ABR25231	Weather Observer	80	99	.32	.39	.35	.34
3ABR27230	Air Traffic Control Operator	60	156	.21	32	.42	.39
3ABR27330	Aircraft Control and Warning Operator	60	133	.30	.38	.66	.55
3AQR90010	Medical Service Fundamentals	60	401	.45	56	.69	.61
3ABR90230	Medical Service Specialist	60	50	.45	.41	.63	.55
3ABR92230	Protective Equipment Specialist	40	60	.53	.26	.42	.37
3ABR63130	Fuel Specialist	40	150	.34	.34	.42	.38
3ABR81130	Security Specialist	40	707	.45	.45	.58	.56

Table 2. **ASVAB Validities for Technical School Courses in The Administrative Cluster**

Course		Minimum AI	N	Mechanical r	Administrative r	General r	Electronics r
				Aptitude Index Validities			
3ABR29130	Communications Center Specialist	60	215	.33	.42	.42	.43
3AQR29222	Printer Systems Operator, Prep	60	91	.14	.34	.35	.34
3ABR29231	Morse Systems Operator	60	84	.01	.28	.25	.22
3ABR29330	Ground Radio Operator, Voice	60	215	.08	.25	.22	.15
3ABR64530	Inventory Management Specialist	60	789	.27	.48	.47	.46
3ABR67133	Disbursement Accounting Specialist	80	122	.11	.32	.21	.22
3ABR73230	Personnel Specialist	60	262	.42	.63	.62	.62

Table 3. **ASVAB Validities for Technical School Courses in the Mechanical Cluster**

Course		Minimum AI	N	Aptitude Index Validities: Mechanical r	Administrative r	General r	Electronics r
3ABR42132	Aircraft Pneudralic Repairman	40	115	.49	.25	.48	.48
3ABR42430	Aircraft Fuel Systems Mechanic	40	66	.13	.37	.43	.45
3ABR43131-A	Aircraft Maintenance Specialist (Reciprocating engine)	40	238	.52	.50	.54	.58
3ABR43131-C	Aircraft Maintenance Specialist (Jet, 1 and 2 engines)	40	691	.38	.34	.40	.37
3ABR43131-E	Aircraft Maintenance Specialist (Jet, over 2 engines)	40	302	.47	.27	.42	.46
3ABR43131-F	Aircraft Maintenance Specialist (Turbo-prop)	40	271	.48	.31	.47	.47
3ABR43230	Jet Engine Mechanic	40	485	.45	.38	.49	.42
3ABR44330	Missile Mechanic	50	53	.50	.46	.60	.64
3ABR46130	Munitions Maintenance Specialist	60	73	.31	.43	.58	.38
3ABR46230	Weapons Mechanic	60	345	.27	.26	.35	.31
3ABR47330	Vehicle Repairman, GP	40	52	.73	.55	.68	.71
3ABR53430	Airframe Repair Specialist	40	150	.55	.46	.47	.56
3ABR53530	Corrosion Control Specialist	50	51	.54	.57	.82	.66
3ABR54330	Electrical Power Production Specialist	50	120	.54	.35	.49	.52
3ABR60531	Air Cargo Specialist	50	170	.43	.36	.48	.42
3ABR60730	Aircraft Loadmaster	50	83	.38	.49	.69	.58

TABLE 4. **ASVAB Validities for Technical School Courses in the Electronics Cluster**

Course		Minimum AI	N	Aptitude Index Validities: Mechanical r	Administrative r	General r	Electronics r
3ABR30130	Aircraft Radio Repairman	80	114	.29	.38	.41	.52
3ABR30131	Aircraft Electronic Navigation Equipment Repairman	80	138	.17	.15	.22	.32
3ABR30133	Electronic Warfare Repairman	80	62	.37	.16	.21	.38
3ABR30134	Aircraft Inertial and Radar Navigator Systems Repairman	80	71	.33	.26	.39	.39
3ABR30430	Radio Relay Equipment Repairman	80	61	.42	.16	.27	.38
3ABR30434	Ground Radio Communications Equipment Repairman	80	70	.44	.25	.33	.43
3ABR30630	Electronic Communications and Cryptographic Equipment Systems Repairman	80	50	.03	.20	.19	.27
3ABR30730	Telecommunications Control Specialist/Attendant	80	82	.41	.41	.42	.43
3ABR32231-A	Weapons Control Systems Mechanic	80	60	.16	.19	.18	.16
3ABR36330	Communications and Relay Center Equipment Repairman, Elec/Mech	60	52	.30	.19	.33	.43
3ABR40230	Aerospace Photographic Systems Repairman	60	66	.26	.20	.27	.27
3ABR42133	Aerospace Ground Equipment Repairman	60	208	.57	.34	.50	.59
3ABR42230	Instrument Repairman	40	68	.32	.30	.36	.39
3ABR42330	Aircraft Electrical Repairman	40	134	.35	.38	.41	.42

TABLE 5. Within Subgroup Sample Sizes

Group	Job Area	Cauc. N	Black N	Other Min. N	Male N	Female N	Total N[a]
01	Intelligence (20X30)	245	[illegible]	–	235	55	290
02	Audiovisual (23X30)	171	43	–	183	31	214
03	Weather (25X3X)	317	55	.	278	96	374
04	Command Control Systems Operator (27X3X)	664	230	–	790	115	905
05	Communications Operations (29130)	369	195	–	[illegible]	158	567
06	Communications-Electronics Systems (30X3X)	1,849	181	53	1,740	343	2,083
07	Missile Electronic Maintenance (31X3X)	544	53	–	517	95	612
08	Avionics Systems (32X3X)	2,163	244	57	2,014	450	2,464
09	Training Devices (34X3X)	170	–	–	158	–	178
10	Wire Communications Systems Maintenance (361/3X0)	226	66	–	303	–	303
11	Wire Communications Systems Maintenance (362X0)	224	69	–	267	–	302
12	Intricate Equipment Maintenance (40X3X)	75	24	–	101	–	103
13	Aircraft Accessory Maintenance (42X3X)	[illegible]	1,041	98	2,187	550	2,737
14	Aircraft Accessory (43130)	193	–	–	177	44	221
15	Aircraft Maintenance (43131)	4,559	1,073	104	4,468	1,268	5,736
16	Aircraft Engineer (4323X)	1,356	363	44	1,431	332	1,763
17	Missile Maintenance (44X3X)	241	52	–	259	36	295
18	Munitions and Weapons Maintenance (46130)	832	162	–	[illegible]	–	1,008
19	Munitions and Weapons Maintenance (46230)	912	154	–	1,084	–	1,084
20	Munitions and Weapons Maintenance (46330)	194	–	–	208	–	209
21	Vehicle Maintenance (47X3X)	251	28	–	282	–	282
22	Computer Systems (51X3X)	[illegible]	–	–	183	86	269
23	Metal Working (53X3X)	[illegible]	160	–	659	168	827
24	Mechanical/Electrical (54X3X)	831	297	–	970	181	1,151
25	Structural/Pavements (55X3X)	505	75	–	471	119	590
26	Sanitation (56330)	215	36	–	251	–	251
27	Fire Protection (57130)	507	188	–	709	–	711
28	Fabric and Rubber Products (58X30)	178	42	–	194	29	223
29	Transportation (60X3X)	1,106	400	40	1,346	200	1,546
30	Food Service (62X3X)	256	136	–	284	117	401
31	Fuel Services (63130)	367	265	–	644	–	646
32	Inventory Management (64530)	1,199	587	83	1,313	556	1,869
33	Material Facilities (64730)	481	360	–	541	317	858
34	Accounting and Finance, and Auditing (67X3X)	439	100	–	372	179	551
35	Administration (70X3X)	1,503	1,078	56	1,716	921	2,637
36	Personnel (73230)	453	180	–	463	183	648
37	Security Police (81130)	2,172	1,222	44	3,438	–	3,438
38	Law Enforcement and Corrections (81230)	1,078	256	–	900	448	1,348
39	Medical (90010)	934	404	28	912	454	1,366
40	Medical (90X3X)	1,385	470	48	1,283	620	1,903
41	Medical (91X3X)	249	48	–	251	49	300
42	Aircrew Protection (92230)	332	63	–	339	63	402
43	Dental (98X3X)	241	68	–	212	108	320

[a] Race N's or Sex N's do not necessarily equal total N. This is because the subsample N's are shown only for subsamples with 24 or more cases on which within subsample validities were computed.

4

Table 6. Educational Index and ASVAB Composite Validities Against Final School Grade

Group	Educ Index	ASVAB Composite AFQT	Mech	Adm	Gen	Elect
01	.38	.42	.25	.30	.40	.37
02	.40	.26	.30	.41	.33	.35
03	.25	.36	.28	.22	.28	.37
04	.22	.38	.28	.14	.35	.34
05	.26	.29	.21	.25	.34	.27
06	.28	.34	.23	.21	.34	.44
07	.30	.37	.22	.23	.31	.45
08	.27	.29	.22	.21	.29	.33
09	.32	.32	.26	.37	.33	.32
10	.23	.26	.29	.23	.31	.32
11	.20	.30	.20	.23	.26	.25
12	.40	.47	.45	.40	.55	.50
13	.26	.31	.40	.18	.31	.36
14	.31	.43	.45	.25	.30	.50
15	.24	.32	.34	.18	.31	.36
16	.32	.42	.43	.33	.40	.46
17	.19	.34	.29	.26	.28	.29
18	.21	.32	.34	.23	.32	.32
19	.22	.37	.27	.22	.34	.37
20	.45	.42	.42	.32	.42	.46
21	.26	.40	.53	.25	.39	.51
22	.13	.32	.05	.27	.26	.24
23	.24	.36	.24	.25	.30	.34
24	.18	.36	.40	.21	.35	.38
26	.16	.24	.34	.12	.17	.26
26	.37	.36	.45	.33	.39	.41
27	.20	.28	.34	.26	.23	.30
28	.28	.28	.41	.19	.09	.25
29	.28	.43	.23	.20	.38	.35
30	.09	.10	.03	.12	-.04	.03
31	.15	.29	.39	.19	.26	.34
32	.27	.30	.18	.13	.34	.29
33	.17	.29	.17	.19	.26	.25
34	.25	.41	.27	.03	.43	.41
35	.23	.32	.16	.20	.32	.27
36	.33	.50	.25	.24	.46	.41
37	.24	.30	.29	.23	.21	.28
38	.30	.38	.32	.26	.39	.39
39	.32	.42	.29	.28	.34	.40
40	.33	.42	.30	.28	.38	.38
41	.31	.35	.21	.25	.37	.30
42	.18	.26	.22	.13	.10	.21
43	.39	.43	.28	.39	.43	.38

TABLE 7. **Validity and Contribution to Prediction of Final School Grade of Educational Background and Test Data**

Group	Predictors[a] (I) Tests + EI	(II) Tests Only	(III) EI Only	F for Contribution of: Tests	EI
01	.54	.47	.38	30.53	29.86
02	.47	.36	.40	7.69	23.76
03	.46	.40	.25	33.80	21.73
04	.42	.40	.22	69.00	10.34
05	.41	.37	.26	32.65	21.77
06	.49	.46	.28	230.17	92.73
07	.50	.48	.30	66.76	20.22
08	.40	.36	.27	126.45	81.38
09	.43	.37	.32	8.66	9.57
10	.40	.34	.23	19.00	15.55
11	.37	.32	.20	16.07	10.47
12	.59	.54	.40	13.59	8.17
13	.46	.44	.26	251.70	54.01
14	.56	.55	.31	35.03	6.03*
15	.43	.42	.24	463.23	106.63
16	.54	.51	.32	242.55	78.11
17	.41	.40	.19	23.40	4.89*
18	.45	.42	.21	98.74	27.64
19	.42	.40	.22	86.99	25.00
20	.55	.48	.45	14.16	20.30
21	.58	.57	.26	56.23	6.75
22	.35	.32	.13	15.18	4.17
23	.41	.38	.24	57.13	27.77
24	.50	.49	.18	162.42	17.19
25	.38	.38	.16	41.07	3.43[b]
26	.54	.49	.37	27.40	19.82
27	.32	.29	.20	23.79	15.15
28	.45	.42	.28	16.63	5.62*
29	.48	.44	.28	153.82	80.37
30	.18	.14	.09	4.70	4.66*
31	.32	.31	.15	28.99	5.00*
32	.38	.32	.27	81.10	96.92
33	.33	.30	.17	36.22	12.98
34	.42	.41	.25	38.19	6.52*
35	.37	.34	.23	129.97	69.56
36	.54	.51	.33	86.72	29.41
37	.36	.31	.24	136.60	136.60
38	.46	.42	.30	102.71	51.91
39	.49	.43	.32	118.79	92.71
40	.50	.45	.33	176.04	124.01
41	.46	.40	.31	20.70	17.10
42	.31	.27	.18	14.70	11.68
43	.54	.49	.39	31.57	24.94

[a]Predictors for the R's in the columns are:
I = AFQT, Selector AI, and Education Index
II = AFQT and Selector AI
III = Education Index only.

[b]Not significant. All other F's are significant at or beyond the .01 level.

*Significant at the .05 but not at the .01 level.

Table 8. **Tests of Hypotheses re Race Equity of Educational Background and Test Data Based Predictions**

	R^a			F for[b]	
Group	I	II	III	H_3	H_{3a}
01	.54	.57	.58	2.09*	.51
02	.47	.51	.52	1.80	
03	.46	.52	.52	4.05**	.02
04	.42	.43	.43	1.95	
05	.41	.42	.42	.86	
06	.49	.50	.50	3.61**	1.39
07	.50	.51	.53	2.48*	1.97
08	.40	.41	.41	5.02**	1.13
09	.43	.47	.48	1.25	
10	.40	.40	.44	1.37	
11	.37	.39	.42	1.62	
12	.59	.60	.62	.77	
13	.46	.46	.46	1.69	
14	.56	.57	.59	1.16	
15	.43	.46	.46	24.15**	3.64**
16	.54	.56	.56	6.00**	1.19
17	.41	.44	.45	1.54	
18	.45	.46	.47	3.08**	1.60
19	.42	.44	.45	3.95**	1.64
20	.55	.58	.61	2.68**	1.53
21	.58	.59	.60	1.16	
22	.35	.35	.38	.94	
23	.41	.45	.48	7.15**	3.77**
24	.50	.50	.50	.89	
25	.38	.42	.42	2.92**	.33
26	.54	.58	.58	1.90	
27	.32	.43	.43	8.88**	.06
28	.45	.47	.50	1.71	
29	.48	.50	.52	10.18**	5.60**
30	.18	.25	.29	2.87**	1.55
31	.32	.42	.44	8.70**	1.59
32	.38	.40	.40	4.71**	2.40*
33	.33	.34	.36	2.91**	2.45*
34	.42	.44	.44	1.30	
35	.37	.39	.40	7.90**	2.02
36	.54	.57	.57	3.89**	1.44
37	.36	.42	.43	28.87**	1.96
38	.46	.49	.50	7.91**	1.03
39	.49	.55	.56	18.85**	2.60*
40	.50	.55	.56	20.42**	10.93**
41	.46	.46	.47	.71	
42	.31	.34	.36	1.67	
43	.54	.56	.57	1.60	

[a]Predictors in the four models are: I = AFQT, Selector AI, Education Index (Problem 3); II = Race, AFQT, Selector AI, Education Index (Problem 8); III = Race, Race x AFQT, Race x Selector AI, Race x Education Index (Problem 9).

[b]H_3 = Knowledge of race contributes nothing to test and EI based prediction of final school grade (Problem 9 vs. Problem 3). H_{3a} = Equation slopes are homogeneous (Problem 9 vs. Problem 8).

*Significant at the .05 level.

**Significant at the .01 level.

TABLE 9. **Tests of Hypotheses re Sex Equity of Educational Background and Test Data Based Predictions**

	R^a			F for[b]	
Group	I	II	III	H_6	H_{6a}
01	.54	.55	.55	.90	
02	.47	.47	.47	.42	
03	.46	.46	.47	1.06	
04	.43	.42	.42	1.06	
05	.41	.41	.41	.27	
06	.49	.50	.50	2.14	
07	.50	.50	.51	1.90	
08	.40	.40	.41	5.43**	6.27**
09	.43	.46	.47	1.92	
11	.37	.39	.40	2.42*	1.26
12	.59	.59	.60	.84	
13	.46	.46	.46	2.08	
14	.56	.57	.59	2.77*	3.42*
15	.43	.43	.44	16.76**	22.37**
16	.54	.54	.55	7.79**	10.14**
17	.41	.41	.45	2.72*	3.53*
20	.55	.56	.58	2.99*	
22	.35	.36	.38	1.87	
23	.41	.42	.43	3.62**	4.35**
24	.50	.50	.51	4.08**	4.82**
25	.38	.38	.41	3.47*	4.32**
27	.32	.32	.32	.92	
28	.45	.45	.49	2.70*	3.44*
29	.48	.48	.48	.90	
30	.18	.30	.31	6.80**	.62
31	.32	.32	.32	.02	
32	.38	.33	.39	1.69	
33	.33	.33	.35	4.68**	5.89**
34	.42	.42	.44	1.99	
35	.37	.38	.38	5.99**	1.64
36	.54	.55	.55	1.35	
38	.45	.47	.47	5.69**	1.04
39	.49	.50	.50	4.92**	2.11
40	.50	.50	.50	4.13**	1.53
41	.46	.46	.48	2.78	
42	.31	.34	.34	1.71	
43	.54	.55	.56	1.98	

[a]Predictors in the four models are: I = Education Index, AFQT, Selector AI (Problem 3); II = Sex, AFQT, Selector AI, Education Index (Problem 14); III = Sex, Sex x AFQT, Sex x Selector AI, Sex x Education Index (PRoblem 15).

[b]H_6 = Knowledge of sex contributes nothing to EI and test based predictions of final school grade (Problem 15 vs. Problem 3). H_{6a} = Equation slopes are homogeneous (Problem 15 vs. Problem 14).

*Significant at the .05 level.

**Significant at the .01 level.

TABLE 10. PREDICTED CRITERION SCORES (ASSUMING MEAN PREDICTOR PERFORMANCE) FOR SELECTED SUBSAMPLES*

COURSE GROUP #	Y' CAUC	Y' BLACK	Y' MALE	Y' FEMALE
04	86.47	84.99	86.05	86.90
05	85.19	83.84	84.68	85.06
06	84.93	82.82	84.91	84.16
08	84.24	81.82	83.93	83.94
13	82.20	81.42	82.04	82.53
15	84.01	80.14	83.21	81.84
16	84.68	81.91	84.09	82.41
18	89.06	87.40	-	-
19	89.30	86.86	-	-
23	84.15	79.19	83.27	82.49
24	85.45	80.32	81.05	79.30
25	-		81.38	80.05
27	86.88	82.96	-	-
29	82.63	79.44	81.95	82.55
30	87.43	84.69	85.42	89.31
31	91.31	87.62	-	-
32	84.32	82.40	83.66	84.14
33	82.96	81.26	82.29	82.83
34	80.46	77.84	80.03	80.00
35	84.04	82.19	82.98	84.03
36	87.07	84.58	86.24	87.07
37	86.00	82.19	-	-
38	83.44	80.59	83.36	81.98
39	83.22	77.94	81.22	82.75
40	82.01	77.39	80.55	81.67
43	-	-	81.33	83.17

* These values are computed only for subsamples with N ≥ 100.

TABLE 11. Average ASVAB Subtest Validities Within School

Aptitude Area	CS	WK	AR	TK	SP	MC	SI	AI	EI
Administrative	.18	.34	.34	-.07	.12	.18	-.01	.01	.06
Electronics	.21	.16	.23	.25	.26	.30	.35	.24	.37
Mechanical	.15	.16	.22	.30	.24	.27	.26	.26	.25
General	.28	.32	.32	.12	.31	.29	.18	.15	.24

TABLE 12. **Validation Samples of Technical Training Courses**

AFSC	Job Title	Selector
27630*	Apr. Aerospace Control & Warning Systems Operator	G**-60
32632	Apr. Integrated Avionic Systems Specialist	E-80
42632*	Apr. Jet Engine Mechanic	M/E-40
43131*	Apr. Aircraft Maintenance Specialist	M/E-50
46130	Apr. Munitions Maintenance Specialist	M/E-60
46230*	Apr. Weapons Mechanic	M/E-60
57130	Apr. Fire Protection Specialist	G-40
60531	Apr. Air Cargo Specialist	A-50
63130	Apr. Fuel Specialist	G/M-40
64530*	Apr. Inventory Management Specialist	A/G-60
64531	Apr. Material Facilities Specialist	A/G-60
70230*	Apr. Administration Specialist	A-40
73230*	Apr. Personnel Specialist	A-60
81130*	Apr. Security Specialist	G-40
81230*	Apr. Law Enforcement Specialist	G-50
90230*	Apr. Medical Services Specialist	G-60

*ASVAB Form 6 samples were available for those AFSCs indicated by an asterisk. ASVAB Form 7 samples were available for all AFSCs listed.

**Composites are M = Mechanical, A = Administrative, G = General, and E = Electronic; M/E indicates either M or E may be the selector.

TABLE 13. **ASVAB-6 High School Composite Validities and Intercorrelations**

AFSC	N	Current Composites: Validity Range	Current Composites: Median Intercorrelation	Proposed Composites: Validity Range	Proposed Composites: Median Intercorrelation
26730	61	-.06/.41	.55	-.18/41	.32
42632	56	.08/.57	.43	.02/.45	.09
43131	341	.32/.48	.62	.16/.46	.32
46230	62	.25/.58	.62	.00/.55	.40
64530	134	.20/.52	.59	-.04/.52	.31
70230	110	.19/.39	.60	.00/.44	.36
73230	54	.09/.50	.58	-.13/.50	.33
81130	400	.25/.40	.62	.07/.42	.37
81230	239	.27/.50	.61	.10/.50	.36
90230	102	.23/.49	.50	-.01/.60	.36

TABLE 14. **ASVAB-7 High School Composite Validities and Intercorrelations**

AFSC	N	Current Composites Validity Range	Current Composites Median Intercorrelation	Proposed Composites Validity Range	Proposed Composites Median Intercorrelation
27630	106	.14/.44	.61	-.01/.44	.27
32632	59	.11/.38	.32	.09/.42	.12
42632	133	.24/.47	.58	.21/.40	.30
43131	623	.22/.36	.49	.09/.36	.20
46130	74	.22/.53	.56	.07/.53	.31
46230	178	.38/.55	.63	.18/.49	.38
57130	113	.21/.38	.51	.09/.41	.21
60531	75	.42/.69	.61	.34/.60	.45
63130	52	.04/.16	.94	.02/.13	.86
64530	251	.23/.47	.60	.05/.47	.39
64531	71	.07/.41	.55	-.05/.41	.31
70230	165	.02/.14	.81	-.08/.16	.66
73230	106	.29/.43	.54	.16/.44	.30
81130	642	.23/.31	.69	.18/.31	.45
81230	359	.24/.41	.68	.09/.41	.52
90230	147	.30/.42	.86	.28/.40	.73

TABLE 15. RELATIONSHIP BETWEEN GEN AI AND GRADE FOR SECURITY POLICE

SKILL LEVEL 5

	N'S		MEANS		
YEARS OF SERVICE	E-4 AND BELOW	E-5	E-4 AND BELOW	E-5	F
04	639	140	52.29	61.07	54.85**
05	234	108	53.35	64.12	41.34**
06	94	104	52.23	58.37	19.83**
07	94	88	54.52	59.77	6.76**
08	63	89	51.59	55.28	3.38
09	35	81	47.43	53.33	8.23**
10	25	54	44.40	53.70	5.86*

SKILL LEVEL 7

	N'S		MEANS		
YEARS OF SERVICE	E-5 AND BELOW	E-6 AND E-7	E-5 AND BELOW	E-6 AND E-7	F
09	414	26	54.54	54.60	.16
10	440	33	53.94	59.85	5.40**
11	360	58	53.53	56.38	2.07
12	357	92	56.33	63.04	13.47**
13	278	104	53.45	60.58	15.87**
14	306	109	50.02	56.61	12.01**
15	351	191	50.33	57.85	21.11**
16	625	396	46.34	56.86	71.17**
17	289	205	44.74	56.71	44.46**
18	345	317	43.80	53.94	42.87**
19	466	586	44.86	56.62	87.97**
20	47	213	50.00	59.44	11.62**

* Significant at the .05 level.
** Significant at the .01 level.

A JOINT SERVICE ACCESSION TEST: PROBLEMS AND PROMISE

Steven Gorman
Headquarters, U. S. Marine Corps
Washington, D. C.

Development of a test battery which has several purposes for several types of users is difficult indeed. Add to this situation the logistical problem of gathering all the test developers and users at a central location, include a tight time constraint, and you have an idea of the environment in which the Armed Services Vocational Aptitude Battery (ASVAB) was developed. The ASVAB currently serves as:

1) a counseling instrument for high schools

2) an accession tool for the four services

3) a differential classification battery for the four services

The users are the thousands of schools testing millions of high school students, and hundreds of Armed Forces Examining and Entrance Stations (AFEES), mini-AFEES, and Mobile Examining Teams (METs), testing millions of service applicants.

The high school composites were redesigned for this present school year to achieve maximum differential classification ability. This was achieved through the use of factor analysis. The resultant composites have high factor loadings and low intercorrelations.

Presently the four services have three metrics for measuring test scores: Navy standard score (Mean = 50, standard deviation = 10), Army standard score (Mean = 100, standard deviation = 20), and percentile scores. Additionally, each service has its own compositing formulas for classification. The common thread among all services is the test score used for determining mental group level, the combination of the word knowledge (WK), arithmetic reasoning (AR), and space perception (SP) subtests. This is referred to as the Armed Forces Qualification Test (AFQT) score, a carryover from pre-ASVAB testing.

One problem which is common to all large scale institutional testing programs is the leakage of test questions/answers to examinees prior to testing. This phenomenon also occurs in the military testing environment. The occurrence of this phenomenon appears to be much more widespread in certain recruiting districts than in others. For managerial purposes, a composite was developed (Sims, 1976) using non-AFQT subtests to predict AFQT and General Technical (GT) scores.

The concept underlying this composite was that the subtests which are vital to enlistment are most apt to be subject to compromise. With this composite, averages of predicted AFQT and GT scores can be compared with the averages of actual AFQT and GT scores. This can be used as a management device to detect significant differences at AFEES, recruiting districts, recruiting stations, and recruiters.

The data sample consisted of 3,081 Marine Corps recruits who were tested at AFEES on ASVAB form 3. Upon arrival at the recruit depots in December 1975 and January 1976, they were administered ASVAB form 6 or 7. They were also administered the Army Classification Battery (ACB-61) as a reference test. To equalize the testing effects, a counterbalanced design was used wherein half the recruits were administered ACB-61 first, and half ASVAB first. The sample was weighted to approximate the normal mobilization population. Multiple regression analysis was conducted to determine the subtests that were the best predictors of the AFQT and GT composites, the two criteria used for accession into the Marine Corps.

Table 1 shows the results of the analysis of raw score AFQT and GT predictors, the standard error of estimate, and the amount of variance accounted for by the predictors. It should be noted that the correlation of the prediction composite with AFQT or GT is comparable to the correlation between alternate AFQT or GT composites.

A likely procedure for reducing test compromise would be to compute the AFQT, GT, predicted AFQT and GT score, and retest those applicants whose predicted and actual scores were statistically different at the .10 confidence level. The retesting would be conducted on an alternate AFQT form, one which is used solely for test score verification. Several test forms are being normed at AFEES presently to be used for test score verification. When properly normed, these AFQT subtests could be used with this procedure to reduce compromise. Another procedure is to collect several months of applicant scores broken down by recruiters and recruiting stations, and determine the statistical

significance of the differences between true and predicted mean score. A management report can then be compiled noting all recruiters and recruiting stations where the mean score differences are statistically significant at the .05 level.

Another problem all services are facing is that of reducing personnel attrition prior to the expiration of active service (Non-EAS attrition). Recent DOD guidelines have established limits on the percent of high school and non high school graduates who will be permitted to be attrited prior to EAS. To reduce this attrition, the services are looking at biographical data and attitudinal items for possible inclusion into a screening test at AFEES.

A study was conducted for the Marine Corps (Sims, 1977) to determine the variables related to non-EAS attrition. As expected, mental ability, age at enlistment, level of education achieved, and number of dependents were all statistically significant variables. Table 4 summarizes the order and value of these variables entering into the regression equation. Tables 3 through 9 present the predicted chances of success for each level of education, age, and mental group. Additionally, however, the study examined the ASVAB subtests to determine if a combination of these could be used to predict attrition.

Employing the data derived from the 3,0[illegible]1 recruits tested in the beginning of 1976 less those recruits who were reservists, a multiple regression analysis was conducted with all ASVAB subtests using the criterion of being in service after 14 months. Table 10 shows the order and value of each subtest entering the multiple regression. The variables which are significant are mostly non-cognitive or speeded subtests, along with educational level achieved. The predictor with the greatest percent of variance explained is numerical operations (NO), a fifty item speeded test. The variable which enters into the regression equation next is the combat scale (CC), a twenty-seven item of the Army Classification Inventory (ACI). The ACI also provides the attrition composite with the attentiveness scale (CA), a twenty item interest test. Another subtest in the attrition composite is a speeded clerical test, attention to detail. Space perception is the only wholly cognitive non-speeded test in the attrition composite. Table 11 summarizes the regression of the ASVAB subtests and education onto non-EAS attrition. Utilizing the attrition composite rather than mental ability significantly increases the attrition prediction. Table 12 is an expectancy table of successful completion of 14 months service for high school graduates, Table 13 for non-high school graduates. Table 14 shows the cross-validation of the traditional and

non-traditional composites, as well as the composites presently used by Navy and Marine Corps.

It is interesting to note the superiority of this non-cognitive composite over the traditional mental ability screen. The difficulty with this composite is that the two interest scales are transparent. For this reason, this composite should be utilized as a recruiting tool, rather than a mandatory selection criterion, since it would quickly lose its utility during periods of recruiting shortfalls.

Another problem with most standardized testing lies in the fact that the same static measurement instrument must be used to assess a wide range of abilities. This procedure increases the testing error from two sources: 1) the test length is necessarily long, which contributes to examinee fatigue; 2) each test will contain questions which are not at an appropriate difficulty level for every examinee: if too easy, the examinee will become bored and may carelessly mark incorrectly; if too difficult, the examinee will guess, thus increasing test noise.

With the recent enhancements to computers and the development of the Owen Bayesian algorithm, testing of personnel abilities can achieve the same precision of ability estimation with a minimized number of items. The Civil Service Commission (Urry,1975) has demonstrated that with an on-line,real-time, adaptive testing sequence, the same precision of ability estimation of a conventional test could be achieved with an average of only one-fifth the number of test items. A computerized adaptive testing (CAT) program at AFEES could provide many benefits. Some of these are:

1) Greater test precision at all ability levels, especially at the tails of the distribution

2) Improved test security

3) Decreased misclassification

4) Reduction of examinee anxiety or boredom

5) Reduction of test length

6) Enhanced applicant motivation with immediate feedback on response results

7) Standardized test administration

8) Improved data quality through elimination of human requirements for calculations and data recording

9) Interface with classification, assignment, and job information systems

The Navy and Marine Corps are working jointly on a CAT project (Gorman, 1977) to demonstrate its effectiveness within a military environment. The project involves psychologists from Headquarters, Marine Corps and the Navy Personnel Research and Development Center, and Marine recruits located at San Diego. All of the research to date is begging the question of not if, but when CAT should be implemented at AFEES. Computerized adaptive testing gives great promise towards more precise and faster ability measurement.

REFERENCES

Gorman, S. Computerized Adaptive Testing with a Military Population, Proceedings of the Computerized Adaptive Testing Conference '77, University of Minnesota, (in press).

Sims, W.H. A Proposed ASVAB Validation Composite, (CNA) 76-3134, Center for Naval Analyses, Arlington, Virginia 15 October 1976.

Sims, W.H. Profile of a Successful Marine, RC 236, Center for Naval Analyses, Arlington, Virginia, 1 July 1976.

Urry, V.W. Computer-Assisted Testing with Livo Examinees: A Rendezvous with Reality, Technical Note 75-3, Personnel Research and Development Center, U.S. Civil Service Commission, Washington, D.C. 1975.

TABLE 1

VERIFICATION COMPOSITE STATISTICS

ASVAB raw scores to be predicted	Best ASVAB prediction	r^2	Standard error of estimate
WK + AR (GT)[a]	GS + MK + GI + MC[a]	.79	5.2
WK + AR + SP (AFQT)[b]	GS + MK + GI + MC[b]	.80	6.2

[a]WK + AR = 2.443 + 0.685 (GS + MK + GI + MC)

[b]WK + AR + SP = 8.50 + 0.849 (GS + MK + GI + MC)

TABLE 2

SUMMARY OF REGRESSION
TO EXPLAIN TOTAL DISCHARGES
(Using conventional variables)

Variable	Cumulative fraction of variance explained (r^2)	Partial F-statistic[a]	Coefficient
ASVAB AFQT	.050	82.5	-.00545
EDUCATION	.066	59.7	-.13726
AGE	.070	11.0	.01747
DEPENDENTS	.070	0.0	
(Constant)			.22857

[a]Partial F-statistic when all variables shown are entered in the regression equation. The critical value is 3.84 for the 95-percent confidence level and 6.63 for the 99-percent confidence level.

TABLE 3

PREDICTED CHANCES OF SUCCESS: PROFILE 1
Age 17

Mental Group	ASVAB AFQT score	Grades of school completed						
		>12	12	GED	11	10	9	≤8
I	93-100	90	90	77	77	72	67	62
II	65-92	88	89	76	76	71	66	60
III A	50-64	84	84	72	72	66	62	56
III B	31-49	80	80	68	68	62	58	52
IV A	21-30	74	74	61	61	56	51	45
IV B	10-20	73	73	61	61	55	51	45
V	0-9	65	65	52	52	47	43	37

TABLE 4

PREDICTED CHANCES OF SUCCESS: PROFILE 1
Age 18

Mental Group	ASVAB AFQT score	Grades of school completed >12	12	GED	11	10	9	≤8
I	93-100	91	91	78	78	73	68	83
II	65-92	89	90	77	77	72	67	61
III A	50-64	85	86	73	73	67	63	57
III B	31-49	81	81	69	69	63	59	53
IV A	21-30	75	75	62	62	57	52	46
IV B	10-20	74	74	62	62	56	52	46
V	0-9	66	66	53	53	48	44	38

TABLE 5

PREDICTED CHANCES OF SUCCESS: PROFILE 1
Age 19

Mental Group	ASVAB AFQT score	Grades of school completed >12	12	GED	11	10	9	≤8
I	93-100	88	88	75	75	70	66	60
II	65-92	86	87	74	74	69	64	58
III A	50-64	82	83	70	70	65	60	54
III B	31-49	78	79	66	66	61	56	50
IV A	21-30	72	72	59	59	54	49	44
IV B	10-20	71	71	59	59	54	49	43
V	0-9	63	63	50	50	45	41	35

TABLE 6

PREDICTED CHANCES OF SUCCESS: PROFILE 1
Age 20

Mental Group	ASVAB AFQT score	Grades of school completed >12	12	GED	11	10	9	≤8
I	93-100	85	85	72	72	67	62	57
II	65-92	83	83	71	71	65	61	55
III A	50-64	79	80	67	67	61	57	51
III B	31-49	75	75	63	63	57	53	47
IV A	21-30	69	69	56	56	51	46	40
IV B	10-20	68	68	55	55	50	45	40
V	0-9	60	60	47	47	42	37	32

TABLE 7

PREDICTED CHANCES OF SUCCESS: PROFILE 1
Age 21

Mental Group	ASVAB AFQT score	Grades of school completed >12	12	GED	11	10	9	≤8
I	93-100	81	82	69	69	64	59	53
II	65-92	80	80	67	67	62	58	52
III A	50-64	76	76	63	63	58	54	48
III B	31-49	72	72	59	59	54	49	44
IV A	21-30	65	65	53	53	47	43	37
IV B	10-20	65	65	52	52	47	42	37
V	0-9	56	57	44	44	39	34	28

TABLE 8

PREDICTED CHANCES OF SUCCESS: PROFILE 1
Age 22

Mental Group	ASVAB AFQT score	Grades of school completed >12	12	GED	11	10	9	≤8
I	93-100	79	80	67	67	62	57	51
II	65-92	78	78	65	65	60	55	50
III A	50-64	74	74	61	61	56	51	46
III B	31-49	70	70	57	57	52	47	42
IV A	21-30	63	63	51	51	45	41	35
IV B	10-20	63	63	50	50	45	40	35
V	0-9	54	55	42	42	37	32	26

TABLE 9

PREDICTED CHANCES OF SUCCESS: PROFILE 1
Age 23

Mental Group	ASVAB AFQT score	Grades of school completed						
		>12	12	GED	11	10	9	≤8
I	93-100	75	76	63	63	58	53	47
II	65-92	74	74	61	61	56	52	46
III A	50-64	70	70	57	57	52	47	42
III B	31-49	66	66	53	53	48	43	38
IV A	21-30	59	59	47	47	41	37	31
IV B	10-20	59	59	46	46	41	36	31
V	0-9	50	51	38	38	33	28	22

TABLE 10

SUMMARY OF REGRESSION
TO SELECT ATTRITION COMPOSITE

Variable	Cumulative fraction of variance explained (r^2)	Partial[a] F-statistic	Coefficient
NUMERICAL OPERATIONS (NO)	.051	18.3	-.00410
COMBAT SCALE (CC)	.074	54.5	-.01453
EDUCATION	.093	45.3	-.11009
SPACE PERCEPTION (SP)	.098	11.1	-.00663
ATTENTION TO DETAIL (AD)	.100	5.9	-.00455
ATTENTIVENESS SCALE (CA)	.101	5.5	-.00597
(Constant)			.89118

[a]Partial F-statistic when all variables shown are entered in the regression equation. The critical value is 3.84 for the 95-percent confidence level and 6.63 for the 99-percent confidence level.

TABLE 11

SUMMARY OF REGRESSION
TO EXPLAIN TOTAL DISCHARGES
(Using attrition composite)

Variable	Cumulative fraction of variance explained (r^2)	Partial[a] F-statistic	Coefficient
ATTRITION COMPOSITE	.086	106.7	-.00476
EDUCATION	.101	52.8	-.12855
AGE	.104	7.9	.01548
ASVAB AFQT	.104	0.0	
DEPENDENTS	.104	0.0	
(Constant)			0.59484

[a]Partial F-statistic when all variables shown are entered in the aggression equation. The critical value is 3.84 for the 95-percent confidence level and 6.63 for the 99-percent confidence level.

TABLE 12

PREDICTED CHANCES OF SUCCESS[a,b]: PROFILE 2

(High school graduates)

Attrition Composite (ATT) raw score	Age						
	17	18	19	20	21	22	23+
180	100	100	100	100	100	100	100
160	100	100	100	98	96	95	93
140	93	92	90	88	87	85	84
120	84	82	80	79	77	76	74
100	74	73	71	69	68	66	65
80	64	61	61	60	58	57	55
60	55	53	52	50	49	47	46

[a]Success probability = 1 - probability of premature discharge

[b]Chances of success = 100 {1-(0.59484-0.00476(ATT)-0.12855+0.0.58(AGE)))

Chances calculated at slightly greater than 100 are reported as 100.

TABLE 13

PREDICTED CHANCES OF SUCCESS[a]: PROFILE 2

(Nonhigh school graduates)

Attrition Composite (ATT) raw score	Age						
	17	18	19	20	21	22	23+
180	99	98	96	94	93	91	90
160	90	88	87	85	83	82	80
140	80	79	77	75	74	72	71
120	71	69	67	66	64	63	61
100	61	60	58	56	55	53	52
80	51	50	48	47	45	44	4[illegible]
60	42	40	39	37	36	34	33

[a]Success probability = 1 - probability of premature discharge

Chances of success = 100(1-(0.59484-0.00476 (ATT)+0.0158(AGE)))

TABLE 14

CROSS-VALIDATION OF PROFILES

Cross-Validation sample	Percentage of variance explained by regression r^2
Profile 1 (ASVAB AFQT, EDUC, AGE)	0.080
Profile 2 (ATT, EDUC, AGE)	0.100
Current Navy (ASVAB AFQT, EDUC, AGE, DEP)	0.080
Current Marine Corps (ASVAB AFQT, ASVAB GT EDUC)	0.077
Original sample	
Profile 1 (ASVAB AFQT, EDUC, AGE)	0.070
Profile 2 (ATT, EDUC, AGE)	0.104

ASVAB: An Adventure in Joint Service Cooperation

by

Major Wayne S. Sellman
Air Force Military Personnel Center

At the 1975 Military Testing Association conference, I served as chairman for a symposium entitled, "Use of a Common Aptitude Test for Entry into All Military Services" (Sellman, 1975). The purpose of that symposium was to provide the background which led to the decision by the Assistant Secretary of Defense (Manpower and Reserve Affairs) to use a common enlistment test and to discuss its development and implementation. Of course, that test was the Armed Services Vocational Aptitude Battery (ASVAB).

Some two years later I am privileged to participate in yet a second ASVAB symposium. This morning's, however, is somewhat different from the previous one in the sense that then we were still four months from implementation and were speaking to "developmental data." Today, we have 22 months experience with ASVAB under our belts and our data are empirical in nature. (Some might say "school of hard knocks" data.)

Well, with all that experience and data now in our possession, how, you might ask, is ASVAB working as a joint service aptitude battery? The answer is - pretty well - especially when you realize that it is intended to meet the selection and classification needs (and, hence, philosophies) of all four services. Undoubtedly, a test designed to serve many masters won't be quite as precise as one working for only one. Yet, in its multifaceted role, ASVAB is a good test (Fischl, Raney, and Seeley, 1978; Swanson, 1978; Valentine, 1977).

As you might imagine, with a test which must be all things to all people, from time-to-time some interesting scientific and management problems arise. To resolve such problems, two joint service committees have been established by the Office of the Assistant Secretary of Defense (Manpower and Reserve Affairs) (OASD/M&RA). The first is the ASVAB Steering Committee. It is chaired by a deputy assistant secretary of defense, and its members

are flag officers from each Service's personnel office. Its charter is to provide policy recommendations on ASVAB development and use to OASD(M&RA). The second, the ASVAB Working Group, consists of testing policy staffers and laboratory scientists from each Service plus representatives from the Military Enlistment Processing Command (MEPCOM). Its responsibility is to "handle" the on-going problems of "building, installing, and maintaining" a joint service test.

The last three years have seen some very interesting developments, both political and scientific, in the ASVAB area. As Chairman of the ASVAB Working Group, I have been intimately involved in almost all of them. Now that ASVAB has become an accepted Service fact-of-life and validity analyses are substantiating its claims of scientific merit, it seems appropriate that the evolution of joint service cooperation be documented. (In short, the behind the scene ASVAB story can be told.) The "ASVAB road" has not always been smooth and certainly not constructed entirely of "yellow bricks" - at least not for the first few miles. Nevertheless, the early days notwithstanding, today ASVAB can be pointed to with pride as an example of a joint service project that worked.

With the above by way of background, I'd like to share with you two examples of joint service interaction/cooperation. They represent, if not the ends of the cooperation continuum, certainly a close approximation thereto. Further, they illustrate the distance traveled in the last three years.

Let's begin with the low end of the scale. In July 1975, the ASVAB Steering Committee met to discuss the status of ASVAB development. One issue that surfaced concerned the inclusion of the Army Classification Inventory (ACI), a short interest test used to select soldiers for combat arms jobs, in the high school version of ASVAB. The Army's position was that whatever version of the test was used in the high schools, it should provide the same scores as the versions used for production testing. The Air Force, Navy, and Marine Corps believed that the high school test with the ACI was too long for use in the schools. With ACI, testing time was three hours, five minutes; without it, two hours, 45 minutes. The Armed Forces Vocational Testing Group (AFVTG) (MEPCOM's predecessor unit, speculated that using ASVAB with ACI they might lose up to approximately 60% of the participating high schools.

Despite all efforts by the ASVAB Working Group, the ACI issue remained unresolved. It again was the main topic of discussion for the ASVAB Steering Committee which met in late August 1975. In addition to the length issue, Air Force and Navy indicated that they believed the ACI was also inappropriate for use in the high schools because of questionable content. Questions had a "weapons, outdoors, sports" orientation which the AFVTG reported might be offensive to some high school counselors and students. Additionally, because it was an interest inventory and not an aptitude measure, it might also present invasion of privacy problems.

Since the Army was the only Service to use the ACI, the Air Force and Navy recommended to OASD(M&RA) that it be deleted from the high school test and administered during the earlier phases of processing at Armed Forces Examining and Entrance Stations (AFEES). The Army retained its position that the ACI was required as input into its classification decisions.

Well, as you can readily see, this is a classic case of the Services agreeing to disagree. After three months further deliberation, OASD(M&RA) finally decided to delete the ACI from the high school ASVAB. The result of this interservice squabble was a six months delay in implementing ASVAB in the high schools.

Now, let's look at joint service cooperation at its best. In February 1977, Professor Lee J. Cronbach of Stanford University wrote the Assistant Secretary of Defense (Manpower and Reserve Affairs) concerning the high school ASVAB. He had been asked by Buros' Mental Measurement Yearbook to review the test and prepare a critique for their next edition. Cronbach had several criticisms--the major one was that the intercorrelations between the high school composites were too high to be of value for vocational guidance. Each Service computes its own set of composites from ASVAB and uses them for classification. In addition, MEPCOM had derived its own composites for use by high school counselors. It was the latter ones taken to task by Cronbach.

As a result of the Cronbach letter, the Deputy Assistant Secretary of Defense (Military Personnel Policy) decided that the high school composites needed to

be reconfigured. In mid-March 1977, he asked the Service personnel R&D laboratories under the auspices of the ASVAB Working Group to develop new ones. By late March, each of the laboratories had developed candidate sets of composites, and in early April the ASVAB Working Group selected the ones proposed by the Army Research Institute. Then, one week later, the ASVAB Working Group assisted MEPCOM in revising the high school counseling materials to reflect the new composites. Because of the concentrated and integrated actions on the part of MEPCOM, the Service personnel R&D laboratories, and the ASVAB Working Group, all revised materials were in the field in time for school year 1977-78. Without joint service cooperation, the entire effort would have been impossible.

Not every ASVAB-related matter that comes along receives the same level of joint service consideration. But, given each Service's unique procurement and placement problems, that is to be expected. The point is that with experience in the joint service arena, trust and rapport between the Services has grown. Now when a problem surfaces, the members of the ASVAB Working Group contact each other - no one operates in a vacuum.

In conclusion, since 1974 we have moved from the concept of a common enlistment eligibility test through the difficulties of its development and implementation to an operational battery administered to over two million examinees annually. I believe that this is a tribute to the dedication, perseverance, and just plain hard work of all those who have been associated with the ASVAB. Its utility has been demonstrated; its support within OASD (M&RA) and the Services is strong. In this healthy environment, I look forward to the next several years of the ASVAB adventure and hope I can take part in still another MTA ASVAB symposium in 1979.

REFERENCES

Fischl, M. A., Raney, J. L., and Seeley, L. C. Preliminary validation of the Armed Services Vocational Aptitude Battery Forms 6 and 7. ARI-TR-78-XX. Alexandria VA: Army Research Institute for the Behavioral and Social Sciences, (in press).

Sellman, W. S. (Chairman) Use of a common aptitude test for entry into all military services. Symposium presented to the 17th Annual Conference of the Military Testing Association, Indianapolis, Indiana, September 1975.

Swanson, C. L. Predictive validation of ASVAB-6/7 in forty-one Navy class A schools. NPRDC-TR-78-XX. San Diego CA: Navy Personnel Research and Development Center, (in press).

Valentine, L. D. Prediction of Air Force technical training success from ASVAB and educational background. AFHRL-TR-77-18. Lackland AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, May 1977.

CODAP in the Design of Concurrent Validity Research

Marvin H. Trattner
U.S. Civil Service Commission

The Civil Service Commission has been engaged for the past three years in a large scale project to assess the criterion related validity of the PACE test. This test is used as part of the examining procedure to fill entry level vacancies in approximately 120 different professional and administrative occupations in many agencies of the Federal government. During the past year the PACE test was administered to 220,000 individuals and 8,000 vacancies were filled from PACE registers. More recent college graduates enter the federal service via the PACE examination than through any other method. Because of the importance of the examination this large scale project was undertaken to assess the criterion-related validity of the PACE test. The criterion related validity for very populous PACE occupations is being determined to supplement the construct validation procedure employed in PACE test development.

The first three occupations studied were Social Security Administration Claims Authorizer, Customs Inspector and Internal Revenue Officer. These are occupations which are unique to government service and for which the PACE is heavily used. Each occupation is found in only one major government agency and has large concentrations of employees in large metropolitan areas.

A concurrent validity paradigm was utilized which assessed the criterion-related validity of PACE for the full performance grade level. The PACE test, criterion instruments, a task inventory and a biographical information blank were administered to currently employed individuals in the grade level which contained the largest number of employees in the occupation. Research participants were tested for eight to twelve hours depending on the occupation studied.

The task inventory was analyzed by the CODAP programs. It can be said to be the keystone of the project and this paper will focus on the procedures employed and the results obtained with the task inventory. This was the first time CODAP was used by the Commission. For this purpose we arranged to install CODAP on a Forest Service UNIVAC computer at Fort Collins, Colorado. Task inventory data for the first occupation studied, the SSA Claims Authorizer, were analyzed for us by our friends at Navy on NOTAP. This was before output from Fort Collins was available. This effort also represented the first major application of CODAP to federal civilian employees outside the Department of Defense.

The task inventory listed the tasks performed in an occupation grouped into major job components or duties. Each task statement consisted of a transitive verb together with an object acted upon. The subject "I" was implicit in the task statements. Respondents indicated whether or not they

perform each task. For all tasks they perform they indicated the relative amount of time they spend performing the task compared to all other tasks they perform. The relative time spent rating was made on a seven-point scale with the following end points -- "very much below average," and "very much above average."

CODAP sums a respondent's relative time ratings and divides each task rating individually by this sum to provide a measure of the relative time spent by the respondent on each task. The time spent in duty performance is the sum of the relative time spent in the tasks which compose the duty. These data are then fed into other CODAP programs which provide very useful analyses. One program calculates a group job description by averaging the individual job descriptions for any specified group of respondents. Another program compares the respondents with each other and clusters them according to the similarity of work performed. Figure 1 in the handout contains the first page of the group job description for Social Security Administration Claims Authorizer.

Research Design

Criteria used in assessing the validity of the PACE consisted of several different measures of job performance. Four criterion instruments were developed: a job information test, a work sample test, a specially devoloped supervisory rating form, and a supervisory ranking form.

Each criterion instrument was scored for the duties composing the job. These were the same duties used to group the tasks included in the inventory. In order to obtain an overall measure of job success for a criterion instrument, the duty scores were weighted by duty importance. The relative time spent in duty performance was used as the measure of duty importance.

The CODAP clustering program was used to indicate the homogeneity of the occupation. This demonstrated whether the research participants were performing the same occupational tasks and consequently could be expected to take the same criterion instruments.

The CODAP group job description listed the most time consuming tasks and duties and was used to check the adequacy of the criterion measures.

Due to time constraints, although the task inventory was developed prior to the criterion instruments, the CODAP output became available only during the final phases of criterion construction. Consequently there was not a perfect degree, but nevertheless a high degree of correspondence between time spent in duty performance and coverage of the criterion measures. For each occupation the task inventory was administered prior to the main data collection to a representative sample of employees. In addition for two of the three occupations the task inventory was administered to all research participants. Also, for Customs

Inspector and Internal Revenue Officer, supervisors rated the occupational tasks for relative difficulty.

In summary, the task inventory served the following functions -

to - test the homogeneity of the occupation,

- check the adequacy of the criterion measures,

- weight the duty scores for a criterion in order to obtain an overall measure of job success,

- select participants who perform in the dominant job type should the occupation prove not to be homogeneous.

Figure 2 in the handout describes the research design. Although there were some variations in the design due to administrative constraints, the figure serves generally to describe the procedure.

Task Inventory Construction

The three occupations studied each appeared in only one major federal agency and were one of the key jobs in those agencies.

In each case we were assured by agency management that the occupation was homogeneous. It was stated that employees at the full performance level (with some minor exceptions) were performing the same tasks and consequently could be administered the same criterion measures. This was alleged to be true regardless of geographical location.

Perhaps because of this situation, the task inventory was constructed with relatively little difficulty. It took an average of 7 subject matter experts working for a week to construct the inventory. We generally employed senior journeymen, working leaders and first level supervisors to write the task statements. The task inventories contained an average of 425 tasks and 11 duties.

Criterion Construction

CODAP data were used differently to construct each of the criteria. The first level supervisor rated and ranked subordinates who participated in the study. Subordinates were rated on a graphic rating scale which described levels of performance on each duty and in some cases on a portion of a duty. It was hoped that focusing attention on duty performance would foster objectivity in supervisory ratings. Consequently more inferential trait ratings were deliberately excluded.

Duty descriptions were written by the subject matter experts who constructed the task inventory. The descriptions included a listing of the tasks judged to be prominent by the subject matter experts. The end points of the scales were similarly defined by them. The ranking form consisted of the duty descriptions with the scale points removed.

The job information test was a multiple choice objective test measuring the examinee's job knowledge. The work sample was a work simulation in which problems similar to those encountered by the journeyman on the job were presented for solution. An attempt was made to make the work sample problems as realistic as possible. Both tests were constructed by subject matter experts employed in the field and also in agency headquarters. Subject matter experts assigned each scorable item to the most appropriate duty so that duty scores could be obtained for each test.

As stated previously, test items were constructed prior to the availability of the CODAP job descriptions. However, when the items were assigned to the most appropriate occupational duty the match between items scored and relative time spent was satisfactory. With one exception the most important duties contained the largest number of test items. That exception was for the revenue officer duty - "locating and contacting tax payer." Moderately difficult test items could not be written for this time consuming duty. Table 1 serves to compare the relative time spent and number of points scored for the job information test and the work sample by occupational duty for the three occupations.

The table shows that the criterion tests measured from 53 to 96% of the job content as determined by relative time spent in duty performance. The rating and ranking forms were developed to record ratings for each duty and consequently they were used to measure performance of the entire job.

It would have been desirable to distribute items to the duties in proportion to time spent in duty performance. However the match is reasonably close and some gain may be achieved by having subject matter experts construct what they consider to be good items without imposing subject matter constraints on them. As previously stated, when overall job performance was scored, the duty scores were weighted by relative time spent.

When the use of a test is challenged in the courts there is a great advantage to being able to document the relevance of criterion measures supporting the use of the test. CODAP output is ideal for this purpose.

For the Claims Authorizer and the Revenue Officer occupations, relative time spent in task and duty performance was used as an indication of importance to occupational success. For the Customs Inspector occupation where law enforcement tasks are performed and are judged to be very important,

relative time spent and relative difficulty level were summed in order to derive a measure of task and duty importance. It was the sum of these two values that was used to weight the duty scores to obtain an overall measure of job success. No ratings of task importance were obtained and hence could not be used for this purpose.

Results

Task Inventory. Table 2 reveals that the agency managers were quite correct. The occupations were very homogeneous as indicated by the average percent overlap for each total group. We have been told by NOTAP personnel that they have never seen group average percent overlap values as high as the ones we obtained. The fact that these were single agency civilian occupations which dealt with subject matter unique to government service and that we only surveyed one grade level contributed to achieving this high homogeneity.

Criterion Instruments

When the criterion duty scores for the various instruments were intercorrelated only some convergent-discriminant validity was obtained. For example, the duty 1 score obtained from the job information test generally correlated no higher with the duty 1 score obtained from the work sample than it did with other duty scores obtained from the work sample. Also the various duty scores obtained for each instrument tended to intercorrelate very highly. The short time limits for the instruments probably precluded reliable differential measurement for the objective criteria. For the Customs Inspector some convergent-discriminant validity was obtained. Procedurally related duty scores for the job information test and the work sample correlated more highly both within and across instruments than did unrelated duty scores. Campbell and Fiske stated in their description of the convergent discriminant validity model that it is very rarely achieved.

Table 3 in the handout describes the internal consistency reliability coefficients obtained for the criterion instruments. It shows very satisfactory internal consistency coefficients for the total weighted scores for the various criteria. These coefficients are similar to ones reported in the literature for similar criterion instruments.

PACE Test. Table 4 contains the validity coefficients for the three occupations. Of the 11 coefficients, 9 are significant beyond the .001 level. Many of the coefficients are very high for job performance criteria and indicate that a great deal of utility will accrue to the use of the PACE test.

The two validity coefficients that were not significant were the supervisory ratings and rankings for Customs Inspectors. The frequent rotation of Customs Inspectors and the independence of their work performance contributed to inadequate supervisory knowledge of the research participants. This may be the cause for the two insignificant correlations.

The results seem to indicate that careful construction of criterion instruments will promote demonstrations of high validity. Usually tests are validated against whatever criteria are conveniently obtainable. Most often the criterion measure is some kind of subjective evaluation made by a supervisor or an instructor. The validities reported using these criteria are less consistent and of lesser magnitude than the ones we obtained.

Few problems were encountered in administering either the task inventory or the criterion measures. No participants complained that the job information test or the work sample contained unfair questions. Some respondents had difficulty comprehending the relative time spent scale. Some objected to reading through the task inventory twice. This was mainly because of the heavy work load we had inflicted on them rather than any specific problem with the task inventory.

No research participants were eliminated from the studies because their CODAP job descriptions differed from the predominant occupational job type. A few could have been eliminated because their job descriptions had low overlap with the major job type, but they were not, due to the high homogeneity values obtained for the total group.

Use of the Task Inventory by Federal Agencies

This was the first exposure to CODAP for the staffs of the three federal agencies. We found some additional uses for the CODAP output in two of the agencies.

In the Social Security Administration the group job description served as the basis for the revision of the Claims Authorizer training course. Similarly the Customs Service plans to employ the output to revise the Customs Inspector course. Also the Customs Inspector task inventory was revised and employed in a new work measurement reporting system. The reporting system describes units of output for Customs Inspectors and will be used to promote efficiency of operations. The Internal Revenue Service has not employed CODAP. This may be because Customs and SSA headquarters managers were more involved in the research.

Conclusions

CODAP can be used very effectively in criterion related validity research to

- determine occupational homogeneity,

- check the adequacy of criterion instruments or to construct criterion instruments,

- weight overall job performance,

- select research participants.

Additionally it can be used to provide excellent documentation for the relevance of criterion measures should the selection test be challenged in the courts.

CODAP in the Design of Concurrent Validity Research

Marvin H. Trattner
U.S. Civil Service Commission

GROUP NUMBER = 1.ORDERED FROM 1 TO 224 CV2327A1 CVISTG11 PAGE 2

JOBDCK REPORT FOR ALL CIVIL SERVICE SOCIAL SECURITY CLAIMS ADJUSTERS

TASK JOB DESCRIPTION-----CASES= 224-----TASKS= 520-----DUTIES= 8-----MEMBERS= 224
SELECTED FROM THE HIERARCHY POSITIONS 1 THROUGH 224----FORMED AT STAGE 1 OF GROUPING ON TIME SIMILARITY.

COUNT OF DUTIES OR TASKS LISTED
CUMULATIVE SUM OF AVERAGE PERCENT TIME SPENT BY ALL MEMBERS
AVERAGE PERCENT TIME SPENT BY ALL MEMBERS
AVERAGE PERCENT TIME SPENT BY MEMBERS PERFORMING
ORDERED BY PERCENT OF MEMBERS PERFORMING

D-TSK	TASK TITLE	% MEMBERS PERFORMING	% TIME SPENT BY MEMBERS PERFORMING	% TIME SPENT BY ALL MEMBERS	CUMULATIVE %	N
A 1	VERIFY THAT ITEMS ON A FORMAL APPLICATION ARE COMPLETE	100.00	0.33	0.33	0.33	
A 42	RECOGNIZE CASES REQUIRING A CHANGE IN BENEFIT AMOUNT DUE ENTITLEMENT TO HIGHER BENEFITS ON ANOTHER A/N	100.00	0.26	0.26	0.59	
B 9	DISTINGUISH BETWEEN SITUATIONS THAT ALLOW OR PRECLUDE THE USE OF MILITARY SERVICE	100.00	0.26	0.26	0.84	
B 30	ESTABLISH IF CHILD IS ATTENDING SCHOOL FULL-TIME	100.00	0.31	0.31	1.15	
B 83	USE COMPUTATION TABLES IN THE ADJUDICATION OF A CLAIM	100.00	0.44	0.44	1.59	5
B 89	ESTABLISH THE DATE OF ENTITLEMENT TO RIB	100.00	0.35	0.35	1.94	
B 90	ESTABLISH THE DATE OF ENTITLEMENT TO AUXILIARY BENEFITS	100.00	0.34	0.34	2.27	
B 91	ESTABLISH THE DATE OF ENTITLEMENT TO SURVIVOR BENEFITS	100.00	0.33	0.33	2.60	
B 94	ESTABLISH THE DATE OF ENTITLEMENT TO HIB	100.00	0.28	0.28	2.88	
D 6	FOLLOW DUAL FOLDER PROCEDURE ON CLAIMS INVOLVING MORE THAN ONE A/N	100.00	0.26	0.26	3.13	10
D 8	ENTER CORRECTIONS ON OA-C101	100.00	0.34	0.34	3.47	
D 10	DIARY A CASE	100.00	0.27	0.27	3.74	
E 4	REQUEST ADDITIONAL EVIDENCE ON SSA-525	100.00	0.38	0.38	4.12	
H 2	VERIFY AND INTERPRET FORM SSA-2, APPLICATION FOR WIFE'S INSURANCE BENEFITS	100.00	0.33	0.33	4.45	
H 4	VERIFY AND INTERPRET FORM SSA-4, APPLICATION FOR INSURANCE FOR CHILD OF LIVING W/E OR S/E PERSON	100.00	0.32	0.32	4.78	15
A 17	EXAMINE REPORTS OF CONTACT	99.55	0.34	0.34	5.12	
A 3	VERIFY THAT APPLICATION IS CONSISTENT WITH OTHER INFORMATION IN THE FILE	99.55	0.34	0.34	5.46	
B 40	ESTABLISH TO WHOM A LUMP SUM DEATH PAYMENT IS PAYABLE	99.55	0.28	0.28	5.74	
B 56	USE THE CLAIMS MANUAL IN THE ADJUDICATION OF A CLAIM	99.55	0.42	0.42	6.16	
C 32	CALCULATE THE ELAPSED YEARS FOR RIB	99.55	0.25	0.25	6.41	20
C 38	CALCULATE THE COMPUTATION YEARS FOR RIB	99.55	0.27	0.27	6.68	
C 44	COMPUTE PIA USING A NEW START METHOD	99.55	0.34	0.34	7.02	
C 51	CALCULATE THE NUMBER OF REDUCTION MONTHS ON A RIB CLAIM	99.55	0.34	0.34	7.36	
C 53	CALCULATE THE NUMBER OF REDUCTION MONTHS ON A WIDOW'S CLAIM	99.55	0.34	0.34	7.70	
D 2	ROUTE CASES TO PROPER COMPONENT WITHIN PC	99.55	0.29	0.29	7.99	25
A 11	IDENTIFY CLAIMANT(S) OR APPLICANT(S)	99.55	0.27	0.27	8.26	
D 13	PREPARE LISTING CODES	99.55	0.22	0.22	8.48	
E 10	REQUEST RELATED FOLDERS TO DETERMINE POTENTIAL OR ACTUAL ENTITLEMENT	99.55	0.24	0.24	8.72	
H 56	COMPLETE OR PREPARE FORM SSA-201, DETERMINATION OF DISALLOWA	99.55	0.29	0.29	9.01	
H 78	COMPLETE OR PREPARE FORM SSA-5015 AND 5015A, DIARY AND LISTING SLIP	99.55	0.29	0.28	9.29	30

Figure 1. Group Job Description for Social Security Administration Claims Authorizer.

PURPOSE	INSTRUMENT	TIME REQUIRED RP	TIME REQUIRED Supervisor
	Job and Demographic Info		
	Task Inventory	2 hrs	2 hrs
	Biographical Info Blank	15 min	
construct criteria and test homogeniety of job; select subjects, weight criteria			
	Criteria		
	Job Information Test	1 hr	
	Work Sample	1 hr & 15 min	
	Supervisory Rating Form		1 hr &
	Supervisory Ranking Form		30 min
	Predictor		
	PACE Test	4 hrs & 15 min	
	TOTAL	8 hrs & 45 min	3 hrs & 30 min

Figure 2. Research Instruments and Development Procedure.

TABLE 1

Comparison of Percent Time Spent in Duty Performance and Percent of Points Scored for the Job Information Test and Work Sample for Three Occupations

	SSA Claims Authorizer			Customs Inspector			Revenue Officer		
	Percent Time Spent	Percent Points Scored JIT	Percent Points Scored WS	Percent Time Spent	Percent Points Scored JIT	Percent Points Scored WS	Percent Time Spent	Percent Points Scored JIT	Percent Points Scored WS
Duty									
1	18	26	9	8	14	8	5	3	
2	35	38	24	3	2		31		
3	29	28	43	27	22	9	12	5	20
4	8	5	24	4	10	35	6	31	14
5	8	2		4	4	19	4	3	2
6	2			22	22	7	2	14	29
7	2			6	4	15	7	10	4
8				6	2	4	8	5	4
9				8	20	2	3	7	
10				6			16	12	27
11				4			1		
12				2			4	5	
13							3	3	
Total percent time spent in task performance measured by criterion test.		96	88		88	85		88	53

TABLE 2

Average Percent Overlap for time spent in Task Performance for Three Occupations

	Average Percent Overlap	N
SSA Claims Authorizer	66	224
Customs Inspector	48	247
Revenue Officer	56	87

TABLE 3

Internal Consistency Reliability Coefficients
for the Criterion Instruments[a]

Criterion	SSA Claims Authorizer	Customs Inspector	Internal Revenue Officer
Job Information Test	.81	.67	.64
Work Sample	.72	.60	.78
Rating Form	.79[b]	.57[b]	.86

[a]Internal consistency for a weighted composite score for all table entries except those noted as b.

[b]Correlation between weighted total rating and ranking score.

TABLE 4

Validity of PACE Test for Three Occupations[b]

	Claims Authorizer	Customs Inspector	Revenue Officer
Work Sample	36[a]	56	56
Job Information Test	61	65	68
Supervisory Rating	30	06	25
Supervisory Ranking	31	03	

[a]All coefficients corrected for unreliability in criterion except Work Sample for Claims Authorizer.

[b]All coefficients significant at $p < .001$ except Supervisory Rating and Supervisory Ranking for Customs Inspector.

A Job Analysis Model for Use In Police Selection

Leon I. Wetrogan, Ph.D.
Cynthia C. Diane
U.S. Civil Service Commission

Introduction

The Personnel Research and Development Center of the Civil Service Commission as part of its mission is responsible for the development and documentation of the entry level examinations for selecting Washington, D.C. Policemen. Approximately two years ago, the decision was made to conduct an extensive job analysis and where necessary develop a new examination procedure for the entry level patrolman job. This decision was based on a number of factors including: 1) the job may have changed since the original research was conducted some years ago; 2) new methods and techniques had been introduced in job analysis since the original research; and 3) new advances in personnel assessment techniques had been developed which might be employed to measure those job related knowledges, skills, abilities and other characteristics, KSAO's, not included in the current examination.

A number of factors were considered in the selection of a job analysis methodology. Among these factors were: 1) the technique would have to allow for the collection of information from a fairly large representative sample of subjects; 2) the method of data collection would have to have face validity in order to obtain and maintain the cooperation of participants; and 3) the method would have to lead to the identification of the important worker KSAO's and lead to the documentation of their linkage to the important or critical job tasks or behaviors.

After a review of the job analysis literature, the decision was made to incorporate the task analysis procedure and a task by ability matching technique to accomplish the job analysis. The major portion of this paper addresses itself to presenting the job analysis model and discussing its use with the Washington, D.C. Police.

Method and Results

The job analysis phase of the project was carried out in four stages.

Stage I: The Identification of Tasks Performed by Police Officers and the Development of a Task Inventory.

In order to generate a comprehensive list of task statements, three sources of information were utilized. They included: 1) a content

analysis of training and operational manuals; 2) a five-day brainstorming session with a panel of seven knowledgeable police officers, representative by race, sex and geographic location in the City; and 3) observation by the present investigators during ride along sessions of patrolman at work.

Next, a panel of four police personnel consisting of one Lieutenant, two Sergeants, and one officer was convened. The panel performed three functions. First, they reviewed each task statement for accuracy and clarity of language. Second, they eliminated task statements that were duplications. Third, they defined the major duty areas represented in the final pool of task statements and sorted the task statements into the duty areas.

The final pool consisted of 317 task statements grouped under fourteen major duty areas. This final pool served as the basis for the development of a Preliminary Patrolman Task Inventory. The Patrolman Task Inventory consisted of a cover letter describing the purpose of the study as well as the purpose of the task inventory. Following the letter were thirteen questions related to demographic, background and experience factors of each person completing the inventory. This information was collected for two reasons. First, the information was used to determine whether the sample included in the study was representative of the D.C. Police force. Second, the data were used to determine if differences on certain demographic variables were associated with differences in the tasks performed.

The remainder of the Inventory contained the instructions for completing the inventory as well as the task statements. The instructions asked each officer to read through the task statements and check those which they had personally performed during the previous twelve months. Next, each officer was instructed to rate each task checked on two scales - relative importance and relative time spent. Data on relative time spent were not collected for the job analysis portion of the project but were included to test some basic research hypotheses. Ratings on relative importance were on a seven point Likert type scale from 1, very much below average in importance, to 7, very much above average in importance.

Once the preliminary Inventory had been developed, it was sent out for review and comment to training academy personnel, top level police department administrators and union officials. Feedback from these sources indicated that the Task Inventory was comprehensive and accurately described the job of patrolman.

Following the review, the preliminary Inventory was administered to a sample of 14 patrolmen, in a pilot study. The sample was representative by race and sex and included two patrolmen from each of the seven police districts. Information collected during the pilot study did not lead to any changes in the Task Inventory, however, it did suggest the need to expand and modify the oral presentation related to the purpose of the research and the instructions for completing the Inventory.

Stage II: Administration of Patrolman Task Inventory

The Patrolman Task Inventory was administered to a sample of 350 patrolmen, fifty from each of the seven districts, during April, 1976. The sample represents approximately 14% of the total D.C. patrolman population. Patrolmen were selected from each district by their supervising sergeant. The only restrictions placed on the selecting sergeants were:

1) the patrolmen had to be in a patrol position rather than a detective position or some other special assignment,

2) the patrolmen had to have a minimum of one year patrol experience and

3) the proportion of males, females, blacks and whites had to be representative of the D.C. police patrolman population.

Stage III: Analysis of Task Inventory Data and Identification of the Most Important Tasks

Of the 350 officers completing the Inventory, data on fifteen were eliminated where: 1) officers were felt to have indiscriminantly marked the task statements because they were all marked with the same numerical rating; 2) officers failed to rate more than 10 percent of the tasks which they had checked as having performed in the previous twelve months; or 3) officers had been performing in a position other than patrolman for more than 25 percent of the previous twelve months. The final sample included in the analyses consisted of 300 males, 31 females, 162 whites and 167 blacks. It should be pointed out that for a very small number of subjects race and sex data were not provided, consequently the sums across race and sex subgroups do not equal the 335 patrolmen included in the total sample.

Separate CODAP Job Description analyses were performed on the Relative Importance and Relative Time Spent dimensions for the total sample, whites, blacks, males and females. Table 1 presents the Average Percent Importance by all Members on the Duties for the total sample, blacks, whites, males and females. Also included in the Table in parentheses are the rank orders for the duties for the total sample and for each sex and race subgroup. The Table indicates perfect agreement across groups on the highest four duties. The most important duty was N, Administrative Activities, followed by H, Conducting Preliminary Investigations, E, Patrolling For Crime Prevention, and L, Conducting an Arrest.

In order to obtain a more precise indication of the agreement on average percent importance by all members for the duties, Pearson Product Moment Correlations were computed between each of the race and sex groups. The correlation for Average Percent importance by all members for duties between males and females was .92. The correlation between blacks and whites for Average Percent importance by all members for duties was .99. These results suggest that there is a high level of agreement between the races and sexes in terms of duty importance.

Since our primary concern involved the identification of the most important tasks, a task level job description analysis was performed. Separate job description analyses were carried out for the total sample, males, females, blacks and whites. Table 2 lists those tasks included in the highest third for the total sample on the Average Percent Importance by all Members. Inspection of the Table indicates that Duty H, Conducting Preliminary Investigations, contained the largest number of tasks in the top third for the total sample, 20. Following Duty H, and tied for second place with 16 tasks each were Duties N, Administrative Activities, E, Patrolling For Crime Prevention and L, Conducting an Arrest. The duty with the fifth highest number of tasks was M, Preparing Cases for Court and Testifying. Duties C, Patrolling for Incidentals, and J, Conducting Follow Up Investigations, had the fewest number of tasks in the top third for the total sample with one each.

Table 3 presents those tasks in the highest third on Average Percent Importance by All Members for each of the race or sex groups which were not in the top third for the total sample. Those task statements in the highest third on the black task level job description are followed by a (B). Those in the highest third for whites are followed by (W). And those task statements in the highest third for females are followed by (F). All of the tasks in the highest third for males were included in the highest third for the total sample. Inspection of Table 3 indicates that there were fifteen tasks in the highest third of the female task level job description that were not in the highest third for the total

Table 1

Average Percent Importance by All Members for the Duties for Total Sample, Males, Females, Blacks and Whites

Duty		Average Percent Importance by All Members				
		Total	Males	Females	Blacks	Whites
N	Administrative activities (supportive)	17.92 (1)	17.93 (1)	17.84 (1)	18.10 (1)	17.76 (1)
H	Conducting preliminary investigations	16.56 (2)	16.55 (2)	16.68 (2)	16.38 (2)	16.69 (2)
E	Patrolling for crime prevention	9.95 (3)	9.98 (3)	9.66 (3)	9.69 (3)	10.27 (3)
L	Conducting an arrest	9.54 (4)	9.58 (4)	9.19 (4)	9.45 (4)	9.65 (4)
M	Preparing cases for court and testifying	8.10 (5)	8.19 (5)	7.24 (6)	8.32 (5)	7.86 (6)
F	Controlling traffic and enforcing traffic laws	7.83 (6)	7.90 (6)	7.18 (7)	7.64 (6)	8.34 (5)
D	Patrolling for community relations	6.53 (7)	6.37 (7)	8.10 (5)	7.11 (7)	6.00 (7)
I	Handling property	4.51 (8)	4.42 (8)	5.42 (8)	4.64 (8)	4.35 (9)
K	Patrolling to apprehend offenders	4.21 (9)	4.25 (9)	3.72 (11)	3.85 (9)	4.54 (8)
G	Caring for the sick or injured	4.05 (10)	4.12 (10)	3.35 (12)	3.80 (10)	4.26 (10)
J	Conducting follow-up investigation	3.63 (11)	3.60 (11)	3.91 (10)	3.68 (11)	3.54 (11)
A	Preparing for tour of duty	3.03 (12)	3.29 (12)	4.07 (9)	3.18 (12)	2.92 (12)
B	Patrolling to determine violations	2.18 (13)	2.20 (13)	1.93 (13)	2.16 (13)	2.18 (13)
C	Patrolling for incidentals	1.85 (14)	1.87 (14)	1.59 (14)	1.90 (14)	1.81 (14)

Values in parentheses are the rank order for each duty within each group.

Table 2

Tasks In The Highest Third On Importance
For The Total Sample

DUTY A - PREPARING FOR TOUR OF DUTY

1. Check proper functioning of radio and siren system
2. Display proper equipment while on duty
3. Load and unload revolver
4. Clean service revolver

DUTY B - PATROLLING TO DETERMINE VIOLATIONS

1. Check for violations
2. Determine ability of occupant/driver to operate vehicle

DUTY C - PATROLLING FOR INCIDENTALS

1. Report fires and accidents

DUTY D - PATROLLING FOR COMMUNITY RELATIONS

1. Use standard automobile equipment
2. Talk to people on beat to establish good relations

DUTY E - PATROLLING FOR CRIME PREVENTION

1. To arrest or prevent the escape of a person who has committed or attempted to commit a crime
2. Transmit and receive on the radio
3. Use standard emergency equipment assigned to vehicle
4. Cruise at low speed while observing for crimes or incidents
5. Check suspicious vehicles for F.I.C.E. (fruits, instrumentalities, contraband and evidence)
6. Check open doors and windows for unlawful entry
7. Use portable radio
8. Check public places while on patrol
9. Separate disorderly person(s) from other persons at scene of disturbance.
10. Separate complaintant from offender in family argument
11. Respond to an emotionally tense crowd condition
12. Restore order after responding to disorderly person call
13. Secure crime scene
14. Protect ambulance crew

Table 2 (cont.)

Tasks In The Highest Third On Importance
For The Total Sample

15. Check inside business establishments to maintain visibility
16. Walk to attain high visibility

DUTY F - CONTROLLING TRAFFIC AND ENFORCING TRAFFIC LAWS

1. Locate and identify witnesses at accident scene
2. Issue traffic violation citation
3. Interview persons involved in and witnesses to a traffic accident

DUTY G - CARING FOR THE SICK OR INJURED

1. Respond to a mentally deranged and dangerous person call
2. Call for ambulance in an emergency
3. Determine injury of person(s) at scene of crime or accident

DUTY H - CONDUCTING PRELIMINARY INVESTIGATIONS

1. Determine if mentally deranged person is dangerous to himself or others
2. Check WALES system for identification of person or property
3. Complete form PD 251: Report on Crime Against Person or Property
4. Locate suspect in crime
5. Determine probable cause to arrest or search
6. Determine the type of violation committed
7. Canvass the surrounding area for stolen car
8. Interview individuals to obtain description of missing person
9. Question suspect before arrest
10. Identify victims and witnesses
11. Evaluate content of interview information obtained from victim or witnesses
12. Interview victim
13. Interview witness
14. Isolate suspect of crime
15. Investigate suspicious persons at scene of crime
16. Visually scan entire building and determine source of break-in
17. Recover all items of evidentiary value at scene of crime
18. Classify incidents to determine the appropriate report
19. Interview complainant concerning crime or incident

Table 2 (cont.)

Tasks In The Highest Third On Importance
For The Total Sample

DUTY I - HANDLING PROPERTY

1. Record information about seized articles on property book
2. Complete form PD 81 (property receipt)
3. Complete form PD 82 (property book) when property is acquired
4. Mark property to be used as evidence for future positive identification

DUTY J - CONDUCTING FOLLOW-UP INVESTIGATION

1. Check with teletype room for repossession or impounding of stolen car

DUTY K - PATROLLING TO APPREHEND OFFENDERS

1. Help secure the safety of an officer in trouble
2. Describe direction of auto to dispatcher when in pursuit
3. Describe vehicle to dispatcher when in pursuit
4. Restrain hostile violators
5. Locate wanted person
6. Pursue suspects on foot

DUTY L - CONDUCTING AN ARREST

1. Complete an arrest for a misdemeanor
2. Advise suspect of rights
3. Prepare form PD 251, Event Report
4. Prepare form PD 255, Arrest Report
5. Prepare form PD 163, Prosecution Report
6. Keep searched prisoner away from others not searched
7. Advise suspect he is under arrest and inform him of the charge
8. Seize the weapon from a suspect
9. Obtain a signed waiver of rights
10. Handcuff a suspect or prisoner
11. Cover front and rear entrances at building where suspect is hiding
12. Place arrested suspects in transport vehicles
13. Arrange for transport vehicles for suspect or prisoner
14. Search for evidence and weapons incidental to a lawful arrest
15. Establish reasonable grounds that subject to be apprehended has committed the crime
16. Search the suspect for fruits, instrumentalities, contraband, and/or evidence

Table 2 (cont.)

Tasks In The Highest Third On Importance
For The Total Sample

DUTY M - PREPARING CASES FOR COURT AND TESTIFYING

1. Report to U.S. Attorney's Office
2. Prepare a traffic case or lesser misdemeanor
3. Prepare Court Papers
4. Produce evidence in court for presentation at trial or hearing
5. Notify witnesses of their scheduled appearance in court
6. Record names and addresses of all witnesses of an incident
7. Testify in felony or serious misdemeanor cases
8. Complete PD 140 (Court Attendance Slip)
9. Relate facts of case to U.S. Attorney or Corporation Counsel
10. Present case to grand jury
11. Testify at preliminary hearing
12. Report to court

DUTY N - ADMINISTRATIVE ACTIVITIES (SUPPORTIVE)

1. Use police communication system
2. Use the call box while on patrol
3. Record run and time on run pad
4. Make proper notifications related to a crime or incident
5. Check all fluid levels in car
6. Check vehicle for damages
7. Check emergency equipment in scout car (lights, siren, etc.)
8. Complete inspection report (PD 7[illegible]) on vehicle
9. Receive and acknowledge assignment from radio dispatcher
10. Call for necessary assistance
11. Aid in training of rookie policemen
12. Obtain the report numbers after a run from the radio dispatcher
13. Go back into service upon completion of a run
14. Book suspect (complete forms 251, 255, 47, 163, 81, 81-A, 82, and PD 68)
15. Operate two-way radio
16. Inform communications branch of the disposition of assignment

Table 3

Tasks In The Highest Third On Importance
For The Race Or Sex Subgroups

DUTY A - PREPARING FOR TOUR OF DUTY

1. Pick up daily hot sheet (F)
2. Inspect crime maps for offense patterns (F)

DUTY B - PATROLLING TO DETERMINE VIOLATIONS

1. Check for permits and their validity (F)

DUTY D - PATROLLING FOR COMMUNITY RELATIONS

1. Assist motorist, in automobile emergencies such as lost keys, stalled auto, flat tire, etc. (B)
2. Establish communications with special interest groups in the community (F)
3. Use map to determine shortest route from one location to another (F)
4. Inform citizens of how to make homes more secure (B)

DUTY F - CONTROLLING TRAFFIC AND ENFORCING TRAFFIC LAWS

1. Use flares at accident scene to prevent further accidents (B)
2. Observe traffic conditions (W)

DUTY H - CONDUCTING PRELIMINARY INVESTIGATIONS

1. Arrange for crime scene search (W)
2. Describe evidence involved in crime in notebook (F)
3. Fill out PD 106 (Flash Lookout) (B)
4. Read broadcast from PD 106 (B)
5. Determine need for additional manpower at a crime scene or unusual incident (F)
6. Identify friends and relatives of missing person for questioning (F)
7. Identify persons entering or leaving crime scene (F)
8. Request owner to report to building following an incident or crime (F)

Table 3 (cont.)

Tasks In The Highest Third On Importance
For The Race Or Sex Subgroups

DUTY I - HANDLING PROPERTY

1. Obtain from claimant positive identification of property such as serial number, distinguishing marks, etc. (B)
2. Place evidence in evidence locker (F)

DUTY J - CONDUCTING FOLLOW-UP INVESTIGATION

1. Maintain communication with people you deal with where a follow-up investigation is necessary (F)
2. Check hot sheet PD 664 for stolen car or missing persons (F)

DUTY K - PATROLLING TO APPREHEND OFFENDERS

1. Chase fleeing suspect with vehicle (W)

DUTY L - CONDUCTING AN ARREST

1. Use physical force to complete arrest (W)

DUTY M - PREPARING CASES FOR COURT AND TESTIFYING

1. Pick up evidence from appropriate clerk for presentation at trial or hearing (B)

DUTY N - ADMINISTRATIVE ACTIVITIES (SUPPORTIVE)

1. Notify shop official and radio dispatcher when radio is malfunctioning (B)
2. Forward to appropriate agency any evidence not processed by or analyzed by MPD (B)
3. Record information from telephone conversations (F)
4. Answer telephone (B)
5. Place prisoner into vehicle (B)
6. Transport prisoner to hospital, court, police station house, central cell block (B)
7. Log final disposition of case on district station house arrest book (F)

sample. The duty with the largest number of tasks was H, Conducting Preliminary Investigations.

Further inspection of Table 3 points out that twelve tasks were in the highest third for blacks which were not in the highest third for the total sample. Of these twelve tasks, five related to Duty N, Administrative Activities. Two each to Duties D and H, and one each to Duties F, I and M. According to Table 3, the task level job description for whites led to the addition of only four tasks. There was one each related to Duties F, H, K and L. Those tasks in the top third for the total sample as well as those in the top third for either race or sex subgroup were retained for further analysis.

Stage IV: The Identification of the Domain of KSAO's and Their Linkup to the Important Tasks

The identification of the important tasks was viewed only as an intermediate step in the job analysis research. The primary goal of the job analysis was to determine the KSAO's necessary for successful performance of the patrolman job. Consequently, the task analysis results were used as the basis for identifying the important KSAO's.

In order to arrive at a preliminary pool of KSAO's, previous job analysis research both related and unrelated to police work were reviewed. Included in the review was the work by McCormick et al with the PAQ, Fleishman with his task taxonomy work, Furcon and Baehr on the Skills Attributes Inventory and Landy and Farr with their Police Performance Description Scales.

Based on the review, 77 KSAO's and their definitions were identified. The KSAO's were next sorted by the present investigators into four broad areas: Cognitive, Social-Personal, Perception, Physical. In the Cognitive area there were 18 abilities including for example, oral communication, number facility, deductive reasoning, creativity, etc. The social personal domain consisted of 24 attributes including tolerance, perserverance, leadership, empathy, etc. The perception area included 16 KSAO's among which were color discrimination, near visual acuity, visual form perception, and size perception. The physical domain consisted of 19 KSAO's including dynamic strength, stamina, multilimb coordination, eye hand coordination, etc. In order to insure the completeness of the KSAO pool and the clarity of the definitions, the list of KSAO's along with their definitions was reviewed by a panel of police personnel. The panel included four patrolmen and one Lieutenant. The review led to some minor modifications in the KSAO definitions but no additions or deletions to the total KSAO list.

In order to reduce the list of KSAO's to a more manageable number, a sample of twenty-one officers, three from each of the seven districts were asked to rate each KSAO in terms of its importance for overall job success. Along with the list of KSAO's each officer was given their definitions. Officers rated the KSAO's on a five point Likert type scale from 1, the ability or personal characteristic is of no importance for successful performance of the police job, to 5, the ability or personal characteristic is extremely important for successful performance of the police job. The mean and standard deviation of the ratings for each KSAO was computed. KSAO's with a mean rating of 3 or higher and a standard deviation less than 1.0 were retained for further analysis. The final list consisted of ten cognitive, twelve social-personal, ten perception and eleven physical KSAO's. Table 4 presents a list of these KSAO's along with their definitions.

The next step in the job analysis process involved linking the forty-three KSAO's to the 137 important tasks and rank-ordering the KSAO's in terms of their importance for successful performance of the tasks. In order to link the KSAO's to the 137 tasks, four ability by task rating forms were developed, one for each of the four KSAO domains. Table 5 presents a sample of one page from the Ability by Task rating form for the cognitive domain. Along with the rating form and instructions, each rater was given a set of definitions for the KSAO's. The instructions asked each officer to begin by getting a clear understanding of the meaning of each ability contained in the Ability by Task rating form. Next they were to begin with task 1 and rate how important each of the abilities was for differentiating superior from barely acceptable performance of each of the 137 tasks. Ratings were obtained on a 5 point Likert type scale from 1, the ability or personal characteristic is of no importance for differentiating superior from barely acceptable performance of the task, to 5, the ability or personal characteristic is extremely important for differentiating superior from barely acceptable performance of the task.

A representative sample, by race and sex, of ten officers from each of the seven districts participated in this phase of the study. Because of the amount of time required to rate the KSAO's in the four domains with the 137 tasks, each officer rated the tasks against only two of the KSAO domains. Consequently, thirty five ratings were obtained on each ability by task combination.

In order to rank order the KSAO's in terms of their overall importance, the mean of the ratings for each KSAO across the 137 tasks was determined. This involved computing the mean rating given by an officer for each KSAO across the 137 task statements. Next, the mean of these

Table 4

Definitions of Skills, Knowledges, Abilities, and Other Characteristics Retained for Task Matching

Cognitive

1. Oral Communication - ability to communicate ideas with spoken words.

2. Deductive Reasoning - ability to apply a broad, general idea or principle effectively to a particular problem or case.

3. Inductive Reasoning - ability to find the most appropriate general concepts or rules which fit sets of data or which explain how a given series of individual items are related to each other. It involves the ability to combine conflicting facts; to logically proceed from individual cases to general principles.

4. Written Communication - ability to write clear and concise letters, reports, descriptions, or instructions.

5. Judgment - ability to solve a problem when all the necessary facts to solve the problem are not given.

6. Following Rules and Procedures - ability to follow rules and procedures in working out job problems.

7. Problem Sensitivity - ability to recognize or identify the existence of problems. It does not include the reasoning related to solving the problems.

8. Problem Solving - ability to find practical ways of dealing with problems and situations.

9. Information Appraisal - ability to evaluate information of an uncertain or conflicting nature.

10. Verbal Comprehension - ability to understand the meaning of words and the ideas associated with them.

Table 4 (cont.)

Definitions of Skills, Knowledges, Abilities, and Other Characteristics Retained for Task Matching

Social-Personal

1. Pressure of Time - ability to work fast and accurately in situations where there is time pressure or emotional strain.

2. Tolerance - ability to put up with and handle verbal abuse from a person or a group.

3. Working to Get Ahead - a liking for work with chances for getting ahead.

4. Leadership - ability to take the lead or take charge when working or dealing with others.

5. Cheerfulness - ability to stay pleasant and good-tempered in dealing with people.

6. Team Work - ability to work as a member of a group

7. Dealing with Attack - willingness to use physical force in dealing in hostile situations.

8. Working Outside - willingness to work outdoors in all kinds of weather.

9. Repetitiveness - ability to perform the same tasks over and over without getting bored or careless.

10. Composure - ability to stay calm and level-headed in difficult, unexpected, or emergency situations.

11. Flexibility - ability to handle unexpected changes on the job, such as new schedules, new routines, or transfers to different jobs.

12. Dealing with People - ability to deal with people politely and helpfully, beyond the giving and receiving of instructions.

Table 4 (cont.)

Definitions of Skills, Knowledges, Abilities, and Other Characteristics Retained for Task Matching

Perception

1. Visualization - the formation of mental images of figures or objects as they will appear after certain changes such as unfolding, rotation or movement of some type.

2. Depth Perception - ability to judge whether objects are near or far away.

3. Near Visual Acuity - ability to see the details of nearby objects clearly (within normal reading distance).

4. Far Visual Acuity - ability to see the details of distant objects clearly (beyond normal reading distance).

5. Visual Form Perception - ability to perceive important detail or configuration in the environment.

6. Closure - ability to mentally organize a disorganized field into a single picture.

7. Night Vision - ability to "see in the dark" or to pick up shapes and movement when lighting is poor or low.

8. Size Perception - ability to estimate about how many objects or people there are in a certain space.

9. Peripheral Vision - ability to see "out of the corner of the eye" when looking straight ahead so as to be aware of things or motion to the side.

10. Sensory Acuity - ability to stay alert over extended periods of time.

Table 4 (cont.)

Definitions of Skills, Knowledges, Abilities, and Other Characteristics Retained for Task Matching

Physical

1. Explosive Strength - ability to expend a maximum amount of energy in one or a series of explosive muscular acts. The ability may be involved in acts such as jumping or sprinting or in throwing objects for a distance.

2. Stamina - ability involves the capacity to maintain physical activity over prolonged periods of time.

3. Static Strength - ability to maintain a high level of muscular exertion for some minimum period of time. This involves the degree of muscular force exerted against a fairly immovable or heavy object in order to lift, push or pull that object.

4. Gross Body Coordination - ability to use the trunk, arms and legs together in movement.

5. Multilimb Coordination - ability to coordinate the movements of two or more limbs (e.g. two legs, two hands, one leg, and one hand). It is most common to tasks where the body is at rest (e.g. seated or standing) while two or more limbs are in motion.

6. Reaction Time - ability to react quickly to signals, unexpected situations, or emergencies.

7. Manual Dexterity - ability to make skillful, coordinated movements of a hand, or of a hand together with its arm. It may involve manipulation of objects (e.g., blocks, pencils), but does not extend to machine or equipment control (e.g., levers, dials).

8. Arm/Hand Positioning - ability to make precise, accurate movements of the hands and arms.

9. Continuous Muscular Control - ability to exert continuous control over external devices through continual use of body limbs.

10. Eye-Hand Coordination - ability to coordinate hand movements with visual stimuli.

11. Rate of Arm Movement - ability to make gross, rapid arm movements.

Table 5

COGNITIVE	ABILITIES											
	Oral Communication	Deductive Reasoning	Inductive Reasoning	Written Communication	Judgment	Following Rules and Procedures	Problem Sensitivity	Problem Solving	Information Appraisal	Verbal Comprehension		
TASKS	A	B	C	D	E	F	G	H	I	J	K	L
1. Pick up daily hot sheet												
2. Transmit and receive on the radio												
3. Display proper equipment while on duty												
4. Inspect crime maps for offense patterns												
5. Load and unload revolver												
6. Clean service revolver												
7. Check for violations												
8. Check suspicious vehicles for F.I.C.E.												
9. Check for permits and their validity												
10. Report fires and accidents												
11. Use standard auto equipment												
12. Use portable radio												
13. Check public places while on patrol												
14. Secure crime scene												
15. Protect ambulance crew												
16. Walk to attain high visibility												
17. Observe traffic conditions												
18. Issue traffic violation citation												
19. Call for ambulance in an emergency												
20. Arrange for crime scene search												
21. Fill out PD 106												
22. Read broadcast from PD 106												
23. Locate suspect in crime												
24. Determine probable cause to arrest or search												
25. Determine the type of violation committed												
26. Canvass the surrounding area for stolen car												
27. Question suspect before arrest												
28. Identify victims and witnesses												

means was computed across raters for each KSAO. This analysis was performed for the total sample as well as for each race and sex subgroup.

Table 6 presents the means for each KSAO for the total sample and for each of the four subgroups. Inspection of the Table indicates that Following Rules and Procedures, Judgment, Oral Communication and Information Appraisal were seen as the most important cognitive KSAO's. Within the Social-Personal domain, Pressure of Time, Repetitiveness, Leadership and Teamwork were viewed as the highest in importance. Looking at the means for the Perception and Physical KSAO's indicates a downward shift in the mean ratings. This suggests that generally KSAO's in these areas are of lesser importance in differentiating successful from barely acceptable performance of the most important job tasks. In the perception domain, Near Visual Acuity, Far Visual Acuity and Visualization obtained the highest mean ratings. For the Physical KSAO's Eye-Hand Coordination, Manual Dexterity and Arm/Hand positioning were rated the highest.

In order to determine if there was a high level of agreement in mean ratings between the racial subgroups and sex subgroups, Pearson Product Moment Correlations were computed across the mean KSAO ratings for blacks and whites and for males and females. The correlation between the black and white mean KSAO ratings was .88 and the correlation between the male and female mean KSAO ratings was .72. These correlations would suggest a high degree of agreement between the races and sexes in terms of mean KSAO importance. Consequently, the development of an examination plan is to be based on the KSAO means for the total sample.

Conclusion

Generally, it is felt that the job analysis model utilized in the present study accomplished our primary goals of identifying those KSAO's necessary for successful performance of the police job and linking them to important job tasks. The procedure was also useful in that it allowed for data collection on a large representative sample and at the same time possesses sufficient face validity to encourage participation.

However, the ultimate test of the job analysis procedure will lie in the validity of the instruments developed from this information about the job. The next steps in the research include: 1) developing an examination plan; 2) developing selection procedures; 3) developing criterion instruments including possibly new supervisory ratings; and finally, 4) conducting a criterion-related validity study including cross-validation.

Table 6

Mean Ratings for the KSAO's for the Total Sample and Race and Sex Subgroups

Skill, Knowledge, Ability, Other Characteristics	Total	Black	White	Male	Female
Cognitive					
Oral Communication	3.08	3.09	3.08	3.03	3.25
Deductive Reasoning	2.67	2.58	2.77	2.65	2.74
Inductive Reasoning	2.59	2.54	2.64	2.56	2.69
Written Communication	2.33	2.33	2.34	2.33	2.37
Judgment	3.15	3.03	3.27	3.05	3.46
Following Rules & Procedures	3.48	3.41	3.55	3.40	3.73
Problem Sensitivity	2.36	2.24	2.50	2.30	2.55
Problem Solving	2.50	2.37	2.64	2.44	2.67
Information Appraisal	2.70	2.46	2.96	2.62	2.95
Verbal Comprehension	2.63	2.52	2.73	2.62	2.63
Social-Personal					
Pressure of time	3.24	3.31	3.13	3.43	2.83
Tolerance	2.46	2.62	2.18	2.67	1.98
Working to Get Ahead	2.78	2.84	2.70	2.95	2.42
Leadership	2.87	2.93	2.78	3.06	2.47
Cheerfulness	2.24	2.27	2.19	2.45	1.79
Teamwork	2.86	2.92	2.75	3.15	2.21
Dealing with Attack	2.21	2.35	1.98	2.37	1.86
Working Outside	2.11	2.18	1.99	2.27	1.77
Repetitiveness	2.86	3.04	2.55	3.12	2.27
Composure	2.73	2.85	2.54	2.90	2.37
Flexibility	2.10	2.16	2.01	2.34	1.58
Dealing with People	2.73	2.86	2.52	2.91	2.35
Perception					
Visualization	1.93	1.91	1.95	1.86	2.14
Depth Perception	1.90	1.72	2.08	1.84	2.07
Near Visual Acuity	2.33	2.05	2.63	2.29	2.47
Far Visual Acuity	1.98	1.87	2.09	1.95	2.06
Visual Form Perception	1.88	1.84	1.92	1.82	2.06
Closure	1.83	1.82	1.85	1.80	1.92
Night Vision	1.88	1.92	1.83	1.83	2.01
Size Perception	1.81	1.82	1.80	1.72	2.11
Peripheral Vision	1.85	1.86	1.85	1.79	2.05
Sensory Acuity	1.87	1.99	1.75	1.72	2.34

Tablo 6 (cont.)

Mean Ratings for the KSAO's for the Total Sample
and Race and Sex Subgroups

Skill, Knowledge, Ability, Other Characteristics	Total	Black	White	Male	Female
Physical					
Explosive Strength	1.54	1.49	1.63	1.66	1.29
Stamina	1.69	1.63	1.77	1.82	1.38
Static Strength	1.51	1.44	1.63	1.64	1.25
Gross Body Coordination	1.72	1.68	1.77	1.87	1.37
Multilimb Coordination	1.95	1.94	1.97	2.03	1.79
Reaction Time	1.92	1.94	1.89	2.00	1.75
Manual Dexterity	2.01	2.04	1.94	2.11	1.77
Arm/Hand Positioning	2.00	2.03	1.95	2.09	1.79
Continuous Muscular Control	1.70	1.69	1.72	1.81	1.45
Eye-Hand Coordination	2.33	2.34	2.32	2.48	2.02
Rate of Arm Movement	1.73	1.73	1.73	1.86	1.43

$r_{\text{Blacks and Whites}} = .88$

$r_{\text{Males and Females}} = .72$

Table 6 (cont.)

Mean Ratings for the KSAO's for the Total Sample and Race and Sex Subgroups

Skill, Knowledge, Ability, Other Characteristics	Total	Black	White	Male	Female
Physical					
Explosive Strength	1.54	1.49	1.63	1.66	1.29
Stamina	1.69	1.63	1.77	1.82	1.38
Static Strength	1.51	1.44	1.63	1.64	1.25
Gross Body Coordination	1.72	1.68	1.77	1.87	1.37
Multilimb Coordination	1.95	1.94	1.97	2.03	1.79
Reaction Time	1.92	1.94	1.89	2.00	1.75
Manual Dexterity	2.01	2.04	1.94	2.11	1.77
Arm/Hand Positioning	2.00	2.03	1.95	2.09	1.79
Continuous Muscular Control	1.70	1.69	1.72	1.81	1.45
Eye-Hand Coordination	2.33	2.34	2.32	2.48	2.02
Rate of Arm Movement	1.73	1.73	1.73	1.86	1.43

$r_{\text{Blacks and Whites}} = .88$

$r_{\text{Males and Females}} = .72$

USE OF CODAP IN TEST DEVELOPMENT

Paul van Rijn

United States Civil Service Commission
Washington, D. C. 20415

Paper presented at
Military Testing Association Convention
San Antonio, Texas
October 1977

Federal and technical guidelines have made job analysis an important requirement in the development of any personnel selection test. Because of this requirement there has been a resurging interest in the evaluation and comparison of various job analysis methodologies for test development. Prominent among the methods being considered is the Comprehensive Occupational Data Analysis Program (CODAP) approach. This approach derives from the research of Dr. Raymond Christal and his associates and was designed primarily to evaluate and develop military training programs and to assist in job classification. CODAP has seen limited use in the civilian section, and, as far as we know, has not yet been used outside the U. S. Civil Service Commission as a primary job analysis method for test development. In the U. S. Civil Service Commission, it has been used to assist in criterion development and research participant selection in the validity research for the Professional and Administrative Career Examination (PACE), and to provide the primary source of task and duty data for entry level police officers and firefighters. This paper generally discusses some of the modifications of CODAP and some of the issues that derive from the use of CODAP in a test development project. Focus will be primarily on the firefighter job.

Despite its name, CODAP represents more than a sophisticated system of computer programs for summarizing job data. It represents the concept of job analysis that job incumbents, rating the relative time spent on tasks they perform, can provide an objective and accurate description of their job. This description can be a useful first step in the development of an examination.

However, the job statement is only as precise as the tasks and duties included in the inventory. Since the inventory developed for a selection test usually focuses on only a single job and must provide information conducive to the identification of the knowledges, skills, abilities and other characteristics (KSAOs) required to successfully acquire and perform the job, there are several important considerations.

First, the task statements must be specific enough to differentiate different ability requirements. If tasks usually occurring together require different abilities, it is desirable to write the tasks as separate statements rather than one single compound statement. This is particularly important when the job, such as firefighting, involves a mixture of diverse abilities, ranging from the cognitive abilities to the physical and sensory-perceptional skills to different personality characteristics, such as interests, attitudes, and motivation.

In developing a task inventory for test development, researchers must be careful that these various aspects of the job are adequately and explicitly reflected among the tasks. Despite extensive reviews by

experts knowledgeable about the job, inventories can still contain distortions. This is not to fault the reviewers, who can generally be relied upon to perform the almost impossible task of determining what tasks are missing from an inventory of several hundred tasks. Rather, the problem is in the volatility of meaning of the English language. An initial firefighting inventory consisting of task statements such as "Maneuver ladders," "Operate hose lines," and "Carry ventilating fans," may appear to be a suitable description of the firefighter job. However, these tasks may not sufficiently reflect the fact that these tasks require KSAOs over and above mere physical strength and dexterity. Reviewers and raters may be well aware that maneuvering ladders and hose requires recall of numerous procedural rules, reasoned interpretation of the specific emergency situation at hand, and considerable judgment in making decisions about where and how best to place the ladder, hose or ventilating fan. However, to assure that these additional aspects of task statements are not overlooked, they are best stated explicitly in separate task statements. For example, "Maneuver ladders," might be supplemented with tasks such as "Determine stability of supporting surfaces," "Understand and follow spoken orders" and "Determine type and size of ladder required."

Since an entry level examination is generally concerned with the abilities required to acquire tasks as well as the abilities required to perform it, it may be similarly advantageous to explicitly include tasks to reflect these requirements. Tasks concerned with reading the manuals, studying the fire department literature, and performance drills help allow these aspects of the job to be more explicitly expressed by the raters.

This specificity and focus of the task statements in a test development effort has the immediate effect that most incumbents will report that they perform a relatively large number of the tasks in the inventory. When this occurs, it should come as no surprise that even the most significant tasks may account for less than one percent of the total ratings and that standardization of raw ratings for each individual rater becomes less important.

The raters themselves are also not typical of respondents to a CODAP inventory. In test development, the raters are the relatively homogeneous group of individuals who perform the job for which the test is being developed. This rater homogeneity, in conjunction with the focus of the inventory, must not be overlooked in the interpretation of the data. Particularly relevant to this is the OVERLAP program.

OVERLAP clusters raters on the basis of the similarities of their ratings. Generally, the differences between major clusters will be of such a magnitude that a different job might be indicated.

However, in test development, the focus of the inventory and the homogeneity of raters is such that clustering may well be on the basis of differences in the response styles of the raters, rather than on actual job differences.

This effect was best illustrated in a study of claims authorizers. Because claims are randomly assigned to authorizers, there must be assumed to be few differences in the claims processed by different authorizers over a year's time. Nevertheless, OVERLAP showed four distinct clusters. Two of these reported performing a very large number of tasks, while the other two reported many fewer tasks.

Interacting with this effect, two groups reported performing the more unusual aspects of claims work, while the other two groups restricted their ratings to the more conventional claims authorizer tasks. Additional follow-up confirmed that there were no job differences and that OVERLAP had in effect clustered raters on their different tendencies to report tasks and on their different tendencies to report unusual tasks.

CODAP provides two estimates of a tasks's relative significance. One averages the ratings only over those who perform the task, the other averages over all raters. The second method is the most useful for test development. Although some extremely critical tasks may not be considered when this approach is used, it must be recalled that a selection test must be designed to measure the generally more significant aspects of the job. It cannot be designed to select on abilities that will be required by only a few workers performing a critical task. The value of a test will be much greater when it selects on abilities required to perform the more moderately critical tasks performed by nearly all workers in the occupation.

Although the relative time spent scale has been demonstrated to be a generally useful indicator of a task's importance, it has certain limitations when a job such as firefighting is considered. For firefighters, the most important or critical tasks are often not the ones involving the most time, and neither of these scales necessarily identifies the tasks that best differentiate among the superior and barely acceptable firefighter. For this reason, the time spent scale was supplemented with a criticality (importance) scale and a difficulty scale. All raters completed all three scales in a counterbalanced design. Time to complete the inventory was two to three hours, and although most raters (69%) indicated that the inventory was too long, many acknowledged the necessity for its length and commented on its comprehensiveness.

Fatigue or changing levels of motivation seemed to have a minimal effect on the ratings, since there were no significant ($p < .05$)

differences among the six sequences in which the 3 scales were administered. Only the difficulty scale showed a significant ($p<.05$) decrease in ratings as the order of administering the scale increased from first to second to third, but this effect accounted for only 4.5 percent of the total variance.

The three scales used contributed uniquely to the job analysis. When the mean ratings (for members performing) on each scale were calculated, criticality correlated only .40 with time spent and .50 with difficulty, while time spent correlated .10 with difficulty. Despite these differences, when only the top quarter of the ranked tasks on each scale are selected using this procedure, 67 were in the top quarter on all three scales, another 34 on two scales. The unique contribution of each scale consisted of four tasks for criticality, five tasks for time spent, and ten tasks for difficulty.

To this list of 120 tasks were added 39 tasks that six supervisors (officers) rated in the top quarter on criticality for the entry level firefighter job. This supervisory perspective of the entry level job contributed mostly to rescue and first aid type tasks as opposed to the emphasis on personal safety and skill acquisition type tasks of the entry level firefighter.

The most difficult and critical phase in the use of the CODAP job description is the link from the task statements to the KSAOs required to acquire and perform the entry level job. Ten representatives from the fire department and four psychologists from the U. S. Civil Service Commission rated the importance of each of 57 KSAOs for acquiring and performing the 159 most significant entry level tasks. To simplify the ratings, raters were asked to indicate if a KSAO was important for predicting performance on the job as a whole, on the duties, and on the individual tasks; and if so, to indicate if the KSAO would be minimally qualifying as a screen-out or if it could be used to differentiate (rank) applicants with different amounts or levels of the KSAO.

Tables 1, 2, and 3 show how psychologists and incumbents ranked the KSAOs in the three categories of abilities used in the study. The similarity between the psychologists' and incumbents' rank ordering of KSAOs is considerable. (Ranks were based on the frequency with which the ability was considered to be a ranking factor.) However, there are some interesting differences. Incumbents considered "Following orders" to be a more important factor than the psychologists, who in turn placed greater emphasis on "Long term memory." Having relatively limited fire scene experience, psychologists may not ordinarily become aware of the extent to which following orders differentiates levels of firefighting performance, while incumbents' lower ratings of long term memory may reflect their

reluctance to dwell on something so basic, all pervasive, and obvious.

The two types of raters showed almost complete agreement in ratings of the perceptual and physical KSAOs (Table 2), but manifested some large differences in their ratings of the miscellaneous KSAOs (Table 3). Of the differences in the overall job ratings of miscellaneous abilities for example, incumbents considered "Willingness to work shifts," "Interest in variety," "Ability to work in confined spaces," and "Ability to drive," relatively more important than the psychologists, who in turn rated "Willingness to work in unpleasant situations" and "Willingness to follow instructions," as more important. Similar differences occur among the duty ratings, although not always in the same direction. Abilities rated highly important for the duties may receive only low ratings on the job overall.

The larger number of differences between raters on these KSAOs may be attributed in part to the less clearly defined nature of these KSAOs and the resultantly lower reliabilities of their ratings. These effects are additionally amplified by the low n (n=4) for the psychologists.

Table 4 shows that psychologists and raters agree considerably in rating the importance of the KSAOs by tasks; however, there are some differences and psychologists do not always link KSAOs to the same tasks as incumbents.

First, psychologists rated basic mechanical ability high, but incumbents rated it low. On the other hand incumbents rated "Ability to use simple formulas" high as a ranking factor where psychologists focused more on basic math as the ranking factor.

Table 5 shows the rater similarity and differences between the tasks linked to the more important cognitive abilities. Included among the larger differences are ratings on "long term memory." Psychologists rated it an important ranking factor for 29 tasks as opposed to 8 tasks for the incumbents. Incumbents focussed on firefighting tasks, psychologists on first aid, special emergencies, and training. For ability 8, "Quick recall," psychologists similarly emphasised the first aid type tasks. Psychologists considered ability 9 and 10, basic math, proportion and percents, important ranking factors for the few firefighting tasks that involve calculations, while incumbents considered "Ability to use formulas" as the critical ranking factor for these abilities.

As indicated earlier, psychologists rated mechanical ability to be important for more tasks than the raters. "Ability to identify problems" was rated highest by both types of raters, although psychologists rated it more important for more tasks, mostly emergency

first aid tasks but also some maintenance tasks. Induction (ability 16) was similarly emphasized by psychologists, although for the first time it included a substantial number of basic firefighting tasks (Duty C). The same is true for deduction (ability 17) but on a much smaller scale.

Judgment (ability 19) was rated high by both groups but on the basis of almost entirely different sets of tasks. Incumbents uniquely linked judgment to tasks involving "recognizing," "determining," and "locating." Psychologists linked it to more specific tasks involving removing or leading persons, safeguarding property and eliminating unsafe conditions.

It is possible to theorize ad infinitum about these differences, but such theorizing should be tempered with the consideration that the psychologist ratings are derived from only four persons and that the criteria for selecting linked tasks to be counted required agreement among 3 of the 4 psychologists as opposed to 5 out of the 10 incumbents. Nevertheless it is important to be aware of some possible differences between psychologists and incumbents' ratings, even though their ratings will generally agree.

For the entry level firefighter project, only the incumbents' ratings were used. Cognitive abilities selected for testing were identified from those that were rated to differentiate performance on a task-by-task basis and on a duty basis.

Both ratings are important. The task ratings link the abilities to their relevant tasks and thereby help to reveal their meaning. Examination of the tasks that require each ability can also help to identify the level at which the ability might be required. The duty ratings allow for an expression of the importance of such abilities as "endurance" which would not be manifested in task ratings alone.

This paper has focused primarily on the identification of ranking factors, especially the cognitive ranking factors. The development of the entry level examination will involve a parallel analysis of those factors that can more properly be used as a screenout in a total examination process.

Table 1

Comparison of Psychologists' and Incumbents' Rankings of the More Important of the Nineteen Cognitive Abilities for Differentiating Firefighter Performance

Ability[a]	Duty Ratings Psychologists		Duty Ratings Incumbents	Overall Job Ratings Psychologists		Overall Job Ratings Incumbents
1. Reading	9		6.5	7		5
2. Writing						
3. Speaking (informal)						
4. Speaking (formal)						
5. Following orders	6	>	1	7	>	2.5
6. Long term memory	1	<	5	3		5
7. Short term memory	7.5					
8. Quick recall	3		3	3		1
9. Basic math						
10. Intermediate math						
11. Math formulae						
12. Advanced math						
13. Basic mechanics						
14. Advanced math						
15. Problem identification	2		4	3		2.5
16. Induction	7.5		9	7		5
17. Deduction	4		6.5	3		7.5
18. Problem solving	10		8	9		10
19. Judgment	5		2	3		7.5

Note. All table entries are ranks based on ratings of four psychologists and ten incumbents of abilities against twelve firefighting duties and the job as a whole.

[a]Complete ability definitions were used throughout the ratings but clarity of presentation precluded their use here.

Table 2

Comparison of Psychologists' and Incumbents' Rankings of the More Important of the Nineteen Physical Abilities for Differentiating Firefighter Performance

Abilities[a]	Duty Ratings: Psychologists	Duty Ratings: Incumbents	Overall Job Ratings: Psychologists	Overall Job Ratings: Incumbents
1. Touch		10		
2. Smell				
3. Hearing		10		
4. Near vision				
5. Color vision				
7. Depth perception				
8. Balance	6.5	10	7.5	
9. Orientation	4.5	7	4	2
10. Watchfulness	2	3	4	3.5
11. Speed estimation				
12. Quick response	1	1	1.5	1
13. Endurance	3.5	2	1.5	3.5
14. Repeated force	4.5	5	4	
15. Short term max-force	8.5	10	7.5	5
16. Max-strength	8.5		7.5	
17. Coordination	3.5	4	7.5	
18. Dexterity	6.5	10		
19. Limb movement		6		

Note. All table entries are ranks based on ratings of four psychologists and ten incumbents of abilities against twelve firefighting duties and the job as a whole.

[a] Complete ability definitions were used throughout the ratings but clarity of presentation precluded their use here.

Table 3

Comparison of Psychologists' and Incumbents' Ranking of the More Important of the Nineteen Miscellaneous Abilities for Differentiating Firefighter Performance

Ability[a]	Duty Ratings		Overall Job Ratings	
	Psychologists	Incumbents	Psychologists	Incumbents
1. Interest in working with people	6	9		
2. Ability to work with People	2	2	4	5
3. Ability to accept "hazing"				
4. Objectivity	7.5			
5. Responsibility and dependability	1	1	2	1
6. Hygiene				
7. Courage	9.5	10.5	2	2
8. Willingness to work in unpleasant situations	5	6	6.5	
9. Ability to get work done on time	3	7		
10. Ability to work under pressure	4	5	2	9.5
11. Willingness to follow instructions		4	6.5	
12. Willingness to work shifts		10.5		5
13. Interest in mechanical objects		8	4	9.5
14. Interest in variety of duties	7.5			9.5
15. Willingness to do routine work				
16. Ability to work in high places				9.5
17. Ability to work in confined places				5
18. Ability to work w/o direct supervision	9.5	3	4	5
19. Ability to drive				5

Note. All table entries are ranks based on ratings of four psychologists and ten incumbents of abilities against twelve firefighting duties and the job as a whole.

[a]Complete ability definitions were used throughout the ratings but clarity of presentation precluded their use here.

Table 4

Comparison of Psychologist and Incumbent Ratings
of Cognitive Abilities,
Showing the Number of Tasks Requiring Each Ability

Cognitive Ability	Number and Rank of Tasks Requiring the Ability		
	Incumbents' Ratings		Psychologist's Ratings
1. Reading	10 (2)	<	7 (7)
2. Writing	0		0
3. Speaking (informal)	2		4
4. Speaking (formal)	2		0
5. Following orders	3		5
6. Long term memory	8 (3)		27 (2.5)
7. Short term memory	1		0
8. Quick recall	5 (5.5)		17 (4)
9. Basic math	2		5
10. Intermediate math	0		5
11. Math formulas	5 (5.5)	>	1
12. Advanced math	0		1
13. Basic mechanics	0	<	10 (5.5)
14. Advanced math	0		0
15. Problem identification	17 (1)		32 (1)
16. Induction	5		27 (2.5)
17. Deduction	1		6
18. Problem solving	4		5
19. Judgment	[illegible] (5.5)		10 (5.5)

Note. Tasks counted are those that at least five out of the ten incumbents or three of the four psychologists linked to the abilities. Tasks are shown in parentheses and ability labels are abbreviated for clarity. The symbols < and > flag some of the larger differences.

Table 5

Comparison of Tasks Derived from Incumbents' and Psychologists' Ratings of Cognitive Abilities for Firefighters

Ability	Tasks Identified by Both Groups	Tasks Identified by the Incumbents alone	Tasks Identified by Psychologists Alone
1. Reading	J12 J28 J14 J30 J15 J31 J16	J20 J26 J21	
2. Writing			
3. Speaking (informal)	K2	A9	B19 K1 K58
4. Speaking (formal)		B22 B26	
5. Following orders	B7 K58	J24	J18 J31 K48
6. Long term memory	CV J12	A6 D29 A13 J26	B1 F1 H1 J14 B10 F3 H7 J15 F7 H8 J16 F9 H11 J18 H12 J20 H13 J21 H14 J23 H17 J31 H18 H21 H22
7. Short term memory		B12	
8. Quick recall	H1 H4 H21	J26 D13	B7 F3 H7 J12 B10 H11 J18 H12 J37 H13 H14 H17 H18 H22
9. Basic math	J12	J25	D27 D28 D29 D30
10. Intermediate math			D27 D28 D29 D30
11. Math formulas		D27 D29 D28 D30 J25	J12 J12
12. Advanced math		J12	
13. Basic mechanics		D36 F5 D39 F10 F1 G18 F2 J12 F3 J20	
14. Advanced mechanics			

(Continued)

Table 5 (Continued)

Comparison of Tasks Derived from Incumbents' and Psychologists' Ratings of Cognitive Abilities for Firefighters

Ability	Tasks Identified by Both Groups	Tasks Identified by the Incumbents alone	Tasks Identified by Psychologists Alone
15. Problem identification	B10 C9 B18	B11 J21	C6 E1 G6 I2 C8 F1 H1 J20 C9 F2 H6 C10 F3 H21 C11 F5 C12 F7 C13 F8 C14 F9 C16
16. Induction	B10 C9 B18	B11 J21	C6 C16 F9 J20 C8 F1 F10 C9 F2 G6 C10 F3 H1 C11 F5 H6 C12 F7 H21 C13 F8 I2 C14
17. Deduction	B10		C3 H1 J42 C5 H18
18. Problem solving	B10 B36 B23 F1		H26
19. Judgment	B10 B	B9 F1 C9 H21	B22 B34 E37 B23 E1 C6 E9 C5 E48

Note. Tasks listed are only those that at least five out of the ten raters and three of the four psychologists linked to the abilities. All tasks correspond to tasks in the Firefighter Tasks Inventory (1975) for the District of

IMPLEMENTATION OF THE CURRENT TASK INVENTORY BANK (CTIB) PROGRAM

by

Gerald R. Clow
Occupational Survey Branch
USAF Occupational Measurement Center
Lackland AFB TX

Introduction: The goal of the CTIB program is to achieve and maintain a current job inventory on every enlisted occupation in the Air Force. This includes, by definition, every career ladder, shred, and special duty identifier (SDI) in existence. The short term goal is to achieve a state of currency on all existing job inventories by April 1978. At last count this was a total of two hundred forty-seven (247) ladders/shreds/SDIs. After April 1978 the goal will be to systematically review existing job inventories and to develop inventories for those occupations that have not been surveyed.

Inventory Development History: Traditionally, it has taken 47 weeks to complete an occupational survey of a single career ladder barring any unforseen circumstances. Until eighteen (18) months ago, eighteen (18) weeks of that time was consumed by the inventory development process. Last year development time was reduced to thirteen (13) weeks. Table 1 contains weekly phase points of the inventory development process which existed from 1967 until 1976. Table 2 shows a reduced inventory development schedule implemented in 1976. As you can see in Table 1, six (6) weeks of development time was consumed by the reproducing and mailing of field reviews.

It was decided in 1976 that the only way to reduce occupational survey time would be to discontinue, on a routine basis, the administration of field reviews. A few specialists and I had long questioned the value of the field review as a significant aid to the development process. In my opinion, field reviews did not provide any significant changes to an inventory which was developed by the interview procedure. The usefulness of the field review was not resolved at this time. However, since most of the inventory development effort was now involved with occupations which had been surveyed, field reviews were judged as not needed. Write-ins from booklets were being kept and were made available to developers. Field reviews would only be performed on a selective basis.

Consequently, field reviews were discontinued as a routine procedure and inventory development time was reduced by five weeks. (One week was added to the six saved because of scanner formatting.) The inventory development process time was thereby established as thirteen (13) weeks.

Thirteen (13) weeks was still unsatisfactory. Projects were being accomplished to meet a goal of performing occupational surveys of fifty-one (51) career ladders a year, and performing a resurvey on each career ladder every four (4) years. By 1976, changes in the classification structure were numerous. User priorities were increasing. It was impossible to schedule fifty-one (51) career ladders a year and be responsive to user demands. Out of this dilemma, the CTIB concept was created.

CTIB Background: In October of 1976, Captain Tom Ulrich and Chief Jim Moon were assigned the responsibility of establishing operational criterion for implementing CTIB. They provided three (3) criteria for evaluating job inventories as stated in Tables 3, 4, and 5. These still serve as a guide for making CTIB judgments.

I initially assigned functional areas, containing a number of occupations, to inventory developers for accomplishment of an initial CTIB review of existing job inventories. This review was to ascertain currency, quality, and priority of inventories. This method did not work. The experience level of some developers was not sufficient to make judgments about quality. It was also counter-productive for developers to shift their attention between CTIB and development work. Managerial control was impossible because of inventory development travel. Therefore, I decided that a separate type of position needed to be created, one that would be responsible for managing a functional area of Air Force occupations for purposes of CTIB.

CTIB Implementation: Three (3) CTIB manager positions were created early this year. Their function is to achieve and maintain a knowledge of a number of occupations. They are responsible for assessing the quality and currency of existing job inventories in their assignment. They must also ascertain user priorities for occupational surveys within their assignment.

CTIB managers accomplish their function by maintaining telephone contacts with training, classification, functional and operational personnel. They also review classification and training documents. However, the most important information to the CTIB manager is through the process of an "interview review" by an inventory developer. This consists of an interview with subject-matter specialists at either a Technical Training Center or a field location whereby the entire job inventory is reviewed in detail. A determination is made after this review as to the quality and currency of a job inventory. Two or three inventory developers are assigned to each CTIB manager, thereby constituting a team. I believe this will, in time, provide a continuity of knowledge about a grouping of occupations that is essential to CTIB.

In Summary, three CTIB managers and their teams are reviewing existing job inventories for quality and currency. This review, and necessary updating of job inventories, should be accomplished by April 1978. At that time, criterion for maintenance of CTIB will be established.

TABLE 1

EIGHTEEN (18) WEEK INVENTORY DEVELOPMENT CYCLE

WEEK	ACTION
1	Research
2	1st TDY (Technical Training Center)
3	Revise Draft
4	Type Draft
5	Field TDY
6	Field TDY
7	Revise Draft
8	Type Field Review
9	Reproduce Field Review
10	Reproduce Field Review
11	Mail Field Review
12	Mail Field Review
13	Mail Field Review
14	Mail Field Review
15	Finalize Job Inventory
16	Finalize Job Inventory
17	Type Job Inventory
18	Send to Printer

TABLE 2

THIRTEEN (13) WEEK INVENTORY DEVELOPMENT CYCLE

WEEK	ACTION
1	Research
2	1st TDY (Technical Training Center)
3	Revise Draft
4	Type Draft
5	Field TDY
6	Field TDY
7	Finalize Job Inventory
8	Finalize Job Inventory
9	Type Job Inventory
10	Type Job Inventory
11	Editorial Review
12	Editorial Review
13	Send to Printer

TABLE 3

CRITERIA FOR PRIORITIZING ASSIGNED CAREER LADDERS

WHEN TO REVISE JOB INVENTORY	WHEN NOT TO REVISE JOB INVENTORY
No Existing OSR	No changes to equipment
Previous inventory incomplete	Previous inventory/OSR complete and comprehensive
Large number of people in career field	Recently revised
Previous inventory did not differentiate between levels or shreds	Small number of people in career field
Recent major revision to CL (AFSC, shreds etc.)	No classification changes to career field or ladder
Tech school, ATC (TT or SG), AFMPC/DPMPRQ or USAF strongly urges revision	Within 6 months of the administration of the inventory
When inventory is not adequate or current	When the field is a direct equivalent for a civilian occupation (such as 522X4, Protective Coating Specialist = Painter) and no change (or only minor changes) has occurred in the methods, materials, or processes in the field since the last inventory
When equipment is added or deleted from the field	
When a process changes (e.g. missile set up on site vs at factory)	
When a change in technology adds or deletes tasks	
When changes to tasks require reorganization of the inventory	
When a large number of tasks are also being performed by another career field or ladder	
Analysis indicates need for change	

TABLE 3 (CONTINUED)

CRITERIA FOR PRIORITIZING ASSIGNED CAREER LADDERS

WHEN TO REVISE JOB INVENTORY	WHEN NOT TO REVISE JOB INVENTORY
No revision for more than ___ years	When information sources (ATC, HQ, MPL) indicate no changes to field anticipated
Multiple ladders in one inventory	When no new tasks have been created for old equipment
	When no new technology has been added to the career field
	When there is no overlap between career field and another career field in the task list

TABLE 4

CRITERIA FOR ESTABLISHING WHETHER OR NOT A JOB INVENTORY IS ADEQUATE AND CURRENT

ITEMS, EVENTS OR DOCUMENTATION WHICH INDICATE A JOB INVENTORY IS ADEQUATE OR CURRENT	ITEMS, EVENTS OR DOCUMENTATION WHICH INDICATED A JOB INVENTORY IS NOT ADEQUATE OR CURRENT
All items in 39-1, STS, POI are included in inventory	Items in 39-1, STS POI are not included
No new equipment or aircraft since last inventory	New equipment or aircraft
No equipment or aircraft deleted since last inventory	Classification or training change since last OSR
Data from last inventory indicates that tasks were broken down to appropriate level	Data from OSR indicates that tasks were not specific enough, data wouldn't provide enough info to make management decisions
Structure of CL and assignment locations have remained constant	Small group survey determines need for revision
No request for update from the field	Proposed career field structure change letters from MPC
Field evaluation by tech school which indicates training is in synch with job requirements	Field evaluation at tech schools indicate training is out of synch with job requirements
A conference of major command career monitors in which no training problems or classification problems are noted	A conference of major command career monitors in which problems concerning training or classification are surfaced
All new equipment tasks are included	New equipment or technology tasks are not included
All new technology tasks are included	
All overlapping tasks are clearly re-stated	Tasks overlap
No tasks are included in inventory that are really done by another AFSC	Another AFSC's tasks are included

TABLE 4 (CONTINUED)

CRITERIA FOR ESTABLISHING WHETHER OR NOT A JOB INVENTORY IS ADEQUATE AND CURRENT

ITEMS, EVENTS OR DOCUMENTATION WHICH INDICATE A JOB INVENTORY IS ADEQUATE OR CURRENT	ITEMS, EVENTS OR DOCUMENTATION WHICH INDICATED A JOB INVENTORY IS NOT ADEQUATE OR CURRENT
Unique tasks performed by that AFSC are included in the inventory	None of the unique tasks performed by that AFSC (if any) are included
Task list is well organized and logical	Task list is disorganized and illogical
Manpower survey tasks are all represented	Previous survey shows over 10% tasks performed by at least 60% of CL
SMS review surfaces no additions or deletions	If none exists and other indicators point to a need to develop one
All data sources agree inventory is adequate (AFMPC, ATC, etc.)	Multiple ladders in one task list
No significant write-ins during administration	Large number of write-ins during administration
Data indicates all tasks performed	Data indicates many tasks not performed

TABLE 5

CRITERIA FOR ESTABLISHING WHETHER OR NOT A CAREER LADDER HAS CHANGED ENOUGH TO WARRANT REVISING THE JOB INVENTORY

EVENTS INDICATING A REVISION IS WARRANTED	EVENTS INDICATING A REVISION IS NOT WARRANTED
New or revised STS, 39-1, add or delete duties or equipment	Current STS and 39-1 are adequately covered in previous inventory
Population of CL has significantly increased or decreased	Population remains stable quantitatively and qualitatively
Assignment locations, major using commands have changed	Assignment locations and major using commands unchanged
Request from Tech School, ATC, and/or HQAF	No classification changes since last survey
Tasks are being added to or deleted from career field as a result of new equipment technology	No changes in course documents for a long time
Unstable career field, has not been surveyed for four years	Stability in terms of equipment, manning and technology
Inventory is disorganized because changes in procedures occurred or new tasks have been added	Same weapons systems remain in field to be worked on as on last survey
New safety standards (e.g. 2 men performing a job formerly performed by a single person).	Survey recently completed
Change in levels of performance on STS	CDC nas remained stable
Change in number of hours in blocks or topics on course chart	STS changes in proficiency levels only
A new AFSC structure is a results of adding or deleting shreoouts, weapons systems or equipment	
Radical changes in methodology of performance	
CDC rewritten	
Change in AFSC structure which creates a new AFSC	
Convention of many authorizations from military to civilian or vice versa	

An Innovation in Identifying

Air Force Qualitative Training Requirements

HENDRICK W. RUCK
MICHAEL W. BIRDLEBOUGH

Under the present Air Force classification, training and assignment policies, first term airmen are normally trained to be universally assignable within their specialty. This need for universally assignable personnel places a requirement on the technical training system to provide broad-based training across the majority of jobs which may be performed by course graduates. Job specific training is then provided by the gaining unit through locally developed OJT programs. This approach to meeting training requirements may result in higher costs than training which is oriented toward a narrower range of particular jobs.

Pressure to reduce resources -- instructors, students, and support -- tied up in training has led the Air Force to reevaluate the necessity for universally assignable airmen. With the prospect of increasing assignment stability, the notion that airmen in their <u>first enlistment</u> should be trained for assignment to any job in their specialty has been questioned. That is, narrowing the training target for first assignments by training to specific jobs or families of jobs

is being examined and tested as a viable alternative to current training and assignment patterns. Such targeting requires a clear picture of current utilization patterns. (The utilization pattern is the number and nature of different jobs within the specialty as well as the relative importance of and relationships among those jobs. Its key dimensions include numbers of personnel assigned, geographic/command distribution, progression, etc.) Occupational survey data are a prime source of information about existing utilization patterns.

A clearly defined outline of the utilization pattern yields payoffs for managers and trainers alike. It allows for the review of jobs performed in the Air Force as a whole as well as within specific commands. Of particular interest are the jobs to which first termers are usually assigned. Such information can be very useful to personnel who manage various specialties, and permits development of well-informed plans for future utilization. Once future utilization has been planned, the training required to support the projected utilization pattern may be developed.

The concept of utilization oriented training requires that training be offered at appropriate career points and using various training modes (such as resident courses, on-the-job training (OJT), and field training) to insure optimum pay-off in training actually applied to mission requirements.

Within each training mode, the full range of available instructional strategies can be employed. The focus of utilization oriented training development within the Air Force has been on first enlistment airmen. However, changes in first enlistment utilization and training require adjustment of utilization and training patterns for the career force as well.

This paper has been quite general to this point, it may now be useful to discuss specific mechanisms which may be used for developing utilization oriented training. Several steps are required in the development of utilization oriented training strategies.

First, the present utilization patterns must be described and understood by decision makers. One proven method for accomplishing this first step is to assemble a small working group consisting of technicians and managers who are responsible for effectively utilizing personnel in the specialty. This group intensively reviews and synthesizes available job data to gain an understanding of present utilization. Although other sources may be used, occupational survey data have effectively served as the baseline for informed decision making. The working group considers the ramifications of that pattern and develops alternative plans for utilization of personnel in the specialty based upon their findings. Alternatives frequently suggested include subdivision of the 3-level

into various channels by weapon system, workcenter, or command of assignment. The results of this working group meeting should then be fully staffed through the personnel, training, and operational agencies to solicit comments as well as considerations not addressed earlier.

Next, a policy-making group comprised of Air Staff, operational, personnel, and training managers as well as technical experts and members of the original working group should meet to finalize the utilization plan. This group has the benefit of the original working group minutes as well as comments from the staffing process. It establishes the utilization plan and develops an action plan for its implementation. Changes to existing personnel and training programs are identified. The results of this meeting are staffed, preliminary cost estimates are developed, and implementing actions are initiated. The actual implementation is an iterative process, with adjustments to the milestone schedule being made as issues arise. Prior to final approval, refined cost estimates are developed which incorporate modifications to the initial plan.

The steps outlined above obviously result in more than just the determination of training requirements. In fact differences between this procedure and the existing system can bring about several significant opportunities. For example, the new procedure allows operational managers to redesign jobs or gain insight into present jobs, which may allow

for less demand on the training system while not seriously affecting accomplishment of the operational mission. The procedure also allows operational managers to review scientifically derived job data for the whole Air Force as well as their own commands -- a capability not normally exercised in the Air Force. Although training requirements are devised based on operational requirements, the procedure allows for operational changes which may provide for more cost effective training.

The procedure described in this paper is not official Air Force policy. It is experimental. It has been applied in three diverse specialties and may be applied to other specialties. Not all specialties are appropriate for such intense consideration. Those specialties which have been analyzed using this procedure have been high training cost specialties. It appears that beneficial changes will occur in training for all three specialties, although final outcomes are difficult to predict. Finally, the procedure is not set in concrete. It has evolved over the course of 18 months and will continue to be adaptable to unique situations.

Small Sample Studies to Assess

Career Field Changes

Capt David S. Street
Capt Douglas Gorman

ABSTRACT

The Air Force Occupational Survey program has for some time been collecting stratified random survey samples which in many cases approach the actual number of cases in a career field population. The current study is planned to assess the relative benefits of administering smaller samples of a career field survey as a first step in developing more quantitative criteria to determine when a career field structure and jobs within a career field might be changing. Since the data analysis is not yet completed, this paper is presented in the form of a description of some of the problems initially encountered as well as some of the issues considered in examining smaller sample surveys for cluster analysis.

Introduction

The Air Force has for several years been collecting occupational data based on survey procedures where a large percentage of the career field population has been sampled. In order to maintain current data on career fields and survey all career fields comprehensively, the USAF Occupational Measurement Center has been surveying approximately fifty-one career fields per year which allows for about one survey per career field every four years. Most jobs in career fields and most career field structures are found to be relatively stable with only minor changes over a four year period. In most cases it seems that the data provided in Occupational Survey Reports (OSR) have proven not to be perishable for at least 3 years and sometimes for much longer periods where a career field remains relatively unchanged. Although this may be a good overall procedure of updating data, it might be possible to more selectively pinpoint needs for data update through the development of decision criteria on data base obsolescence. The use of the small sample study is one step in developing some idea of the feasability of such a technique.

Problem: Change in Career Field Data

Career fields can change in two primary ways: (1) by adding or deleting tasks performed or (2) by changing the amount of time different people expend in given tasks. Both of these may constitute changes in career field structure in relation to Air Force management procedures. With the advent of the Current Task Inventory Bank (CTIB) procedure at the USAF Occupational Measurement Center it has become possible to update task inventories systematically on a realtime basis. What this means in terms of career

field change is that tasks may now be added to the survey instruments or identified as possibly no longer performed. This gives the potential of having any inventory current for administration at any point in time. Once "significant" changes have occurred in the tasks or in who performs tasks in a career field or between career fields, then the OSR data might require updating. If the change is reasonably minor, it might be possible to "correct" the current data on the basis of a more limited sampling technique. If such a procedure could be developed, then it might be possible that a system more sensitive to career field change could be adopted to update occupational data in the Air Force.

Method of Approach

In examining the potential of a more limited sampling size for updating survey data one might initially suspect that a statistical solution could be calculated to determine what would be "scientifically" acceptable. However, when one examines the clustering analysis technique [1] which is so useful in this type of job analysis, then it becomes clear that although a statistical solution might be developed, it will probably be complex. This technique might be thought of graphically as taking in more and more bits of information in a photograph or a jigsaw puzzle, the more information you have, the more recognizable the picture becomes. The computer with the Comprehensive Occupational Data Analysis Programs (CODAP) will cluster the job incumbents together, but it is up to the occupational analyst to be able to see the picture and interpret the "picture" and meaning to the manager who will use the data. Therefore, although some criteria might be arranged for the decision of meaningful job groups, the analyst is placed in the position of re-evaluating the criteria as they might be

applied to any given career field, set of career fields, or possible data utilization.

In approaching the problem of determining when a career field might be "significantly" changing, two elements of interest would be the tasks added or deleted and what may be called the structure of the career field which relates to the relative amount of time spent by job incumbents performing any given task. Task additions or deletions might alter a task inventory to an extent, but in terms of perishability of the data the real concern is how the career field is organized into job type or cluster groups and what tasks these groups may be said to perform most often. Also a career field might change significantly through a management decision where the same tasks were redistributed in terms of groups performing a set of tasks. Obviously, any one of these kinds of changes might have an impact or training, classification, or assignment practices and is vital information to effective management. A measure of change in career field "structure" would be of primary interest. Career field structure or job structure might be defined as the number and relationships of jobs in a given career field and the number of job incumbents related to any given grouping.

The Air Force Human Resources Laboratory's CODAP system prints out an occupational structure in the form of a cluster merger diagram along with a series of computer arrayed data products which allow an occupational analyst to interpret the data received in an Air Force Occupational Survey. Normally, these programs are used to compare job types within a career field or between career fields when two or more career fields have been surveyed in the same survey instrument. However, the CODAP system is a highly flexible series of programs and the small sampling project might use

it to advantage in systematically comparing different job structures across several different samples.

The question of "significant" change in relation to job structure must always be considrered within the context of the relative management policies and concerns of the job environment. Many items may be considered in terms of significance such as worker satisfaction, mission effectiveness, training costs, classification limitations, assignment system adaptability, recruiting policies, and a host of other possibilities. Here significant change is tied back into a more socially defined term that may relate to the purpose that the user might make of the data. It seems safe to assume that no simple definition of significance will suffice in this type of data feedback environment. A change in any portion of the behavioral "picture" of a career field might or might not effect the management perception of the whole, and the viewpoint of "significance" might be judged in relation to the aspects individually as well as synergistically.

Focusing on a measure of significant change in career field structure would seem, at least initially, to require a significant amount of qualitative comparison in order to tease out as many relevant aspects of the situation as possible. The question of whether or not a career field structure has changed must always be defined ultimately in relation to who might be concerned. It might be that for certain data users broad information with significant detail would be required and a more comprehensive and updated survey might be necessary. For another data user, a limited set of job groups might be of interest. For still another data user general information might be all that was required and whatever changes that might be occurring might not be relevant to that level of management questioning. Each request might require a different level of need for data currency. Thus a decision to resurvey a field could be approached from

several viewpoints. If a small sample could be relied upon as being sensitive to a certain degree of differentiation, existing data might be revalidated for even detailed questions or issues.

The initial question might then be posed as "How sensitive is a smaller, more limited survey sample, in reflecting the structure of a career field?" Another way of looking at this question would be to ask "How much information and what kind of information is lost in relation to career field structure when a sample size is diminshed?" In beginning to develop techniques for dealing with how often a survey would need re-administration, and if smaller than 100 percent samples are to be used, this first step of evaluating the sensitivity of sample size seems like a useful starting point.

Plan of Research

In order to investigate possible information loss in smaller samples a gradated series of occupational analyses will be performed. Initially one data set was selected from available career fields currently being analyzed. From this data file four separate random samples will be drawn (75%, 50%, 25% and 10% of the total original survey cases). Each of these samples will then be run as an independent cluster analysis by the CODAP system. Since the randomized sampling will be drawn independently from the total survey case file, there should be some degree of repetition of cases within the four samples. After all the studies have been job typed, comparisons of the clusters across the four samples and with the original survey clustering structure will be made. Comparisons will then be made between the job groups among the four subsample populations to determine the degree of similarity between job groups in different survey samples.

If similar job structures emerge in the cluster diagrams multiple comparisons will be made of what groups are not represented in the smaller samples as well as some judgement of what information might be made less evident by moving toward a smaller sample.

The Air Traffic Control Radar Repair Career Ladder (AFSC 303X1) Survey [2] was selected as a data file because it has just recently been surveyed and appears to be a relatively stabile career field. The useable survey returns included 1,111 cases and represent 56% of the career ladder. Twenty different job groups were defined in the last survey report indicating a fairly heterogeneous career ladder. The reported group sizes from this OSR will be used to measure the representativeness of the four random sample groups.

Discussion

The analysis of the four studies will be carried out by one occupational analyst. As an initial study much of the procedure will involve relative comparison for the heuristic value of determining if some pattern or trend in the data might exist. Squadron Leader William J. Watson (RAAF) completed a similar type of study while working at AFHRL in 1974. [3] His study used a sample of 1,983 cases drawn at random from 5,544 cases collected. This analysis was compared to an analysis of a second set of data drawn randomly from the 3,561 remaining cases. All the job types identified in the first analysis were also identified in the second with the exception of six small job groups accounting for 73 cases. This study demonstrated that equal samples drawn from the same population without substitution would yield a roughly similar analysis. By taking this subsampling approach to a smaller sample size and sampling with replacement after each

percentage draw it is hoped that the sensitivity of the clustering technique with small samples can be assessed. Considering this previous study, it might be expected that some job groups might be "missing" in any one of the samples without being interpreted as being less accurate or representative. The question of whether or not there would be a "significant" difference would depend on whether or not an Air Force manager might come to identical conclusions based on data from the various samples. Any set of decision criteria later to be developed from the initial findings of this study would require consideration of the realm of data utilization.

REFERENCES

1. Ward, J.H., Jr. "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, 1963, 58, pp. 236-244.

2. Eustis, James N. and Street, David S. Occupational Survey Report of Air Traffic Control Radar Repair Career Ladder (AFSC 303X1), AFPT 90-303-198, Lackland AFB, TX, USAF Occupational Measurement Center, 15 May 1977.

3. Watson, William J. The Similarity of Job Types Reported From Two Independent Analyses of Occupational Data. Lackland AFB, TX, Air Force Human Resources Laboratory, AFHRL-TR-73-58, February 1974.

AN INVESTIGATION OF FIVE ALTERNATIVE INSTRUCTIONAL STRATEGIES

Charles W. Howard, Ph.D

INTRODUCTION

The field of instructional technology includes research and development in the areas of programmed instruction, audio visual equipment and digital computers. Each area has developed instructional materials or software to improve the quality of instruction. One important problem is how to modify existing instructional programs which have not been developed utilizing a systems approach. In addition to insuring post-instructional competence of the trainee, instructional strategies must also be directed toward the transfer of learned skills to applications of those skills. A fundamental problem shared by instructional technologists is modifying or developing materials from existing content materials. Each area has developed instructional materials specified by objectives and evaluated by objective based tests to insure the quality of instruction. Instructional technologists utilize the Instructional Systems Development (ISD) approach for analyzing the performances to be learned and designing instructional systems to insure that the desired instructional outcomes are realized.

The Department of Army has begun converting its training program into self-instructional sets of materials. Numerous programs are currently self-instructional, however, many instructor-taught programs are still used and in need of adaptation. James E. Briggs (1964), Smith (1966) identified two potential problems in converting existing lecture/demonstration course materials into self-instructional programs which include 1) time to produce effective sets of materials; and 2) available programmers. Because of time constraints and a lack of available programmers, alternate strategies for converting typical course materials need to be analyzed.

With this mission, it is necessary to examine (a) the effects of alternative strategies for instructional development of self-instructional materials, and (b) whether instructional developers can modify existing software into self-instructional programs which promote learning by employing auto-elucidation techniques.

The primary purpose of this study was to determine the efficiency of five types of instructional development strategies. Efficiency was investigated in terms of achievement and time. The secondary purposes were: 1) to investigate the utility of informational feedback for each instructional development strategy, 2) to investigate the utility of adjunct programming in the form of post-text questions and remedial branching.

The five instructional strategies that were explored in this study were:

1. instructional text;

2. instructional text supplemented by post-text questions;

3. instructional text supplemented by post-text questions along with knowledge of results of the post-text questions;

4. instructional text supplemented by post-text questions, knowledge of post-text questions and directions for remedial instruction;

5. programmed instruction, developed using the identical content of the instructional text and containing linear and branching programming.

The independent variable in this study was instructional strategy. The dependent variables in this study were adjusted post-test scores, post-test time, total package time and total teaching time.

METHODOLOGY

Sample

The population from which the sample was drawn consisted of persons serving as enlisted or commissioned engineering personnel for the United States Army. This study was conducted at Fort Belvoir Army Engineering School, Belvoir, Va.

The sample consisted of 720 subjects who were assigned to various technical training programs by the Department of the Army. The subjects were randomly assigned to one of the six groups, i.e., five treatment groups and one control group. This random assignment yielded 120 subjects per group.

Pre-test scores were used to determine the level of mastery of each subject. Each subject was classified as a non-master and therefore admissable as a subject for this study.

Treatment

The five instructional stragegies consisted of content material, questions, knowledge of results and remediation appropriate to the strategy.

The primary component was the content material. The content material used for this study described the identification and functions of components for a low-voltage circuit tester.

The next component of the instructional strategy was the issuance of statements or questions. The questions were derived from the content material and were designed to facilitate learning for knowledge and comprehension of the content material.

Knowledge of results (KR) was the next component. This component consisted of subject matter experts responses to the questions. For the purposes of the study, distinctions between instructional strategies were made by providing KR to the learner in three of the five instructional strategies. The inclusion of KR was a form of feedback, however the implied use of KR was limited to informational feedback. The KR component was designed to provide the learner with the model performance expected and therefore enable the learner to make a comparison of the model performance and his actual performance.

The final component of the instructional strategy was remediation. This component consisted of a written statement directing the subject to appropriate portions of the content material when actual performance made by the subject was not in agreement with the model performance.

Design

The research design proposed to test the hypotheses for this study consisted of an analysis of variance model and an analysis of co-variance model. The ANOVA model was designed to separately test the dependent variables: post-test time, total package time, and teaching time. The ANCOVA model was designed to test the dependent variable post-test score adjusted for the influence of the pre-test score. The independent variable was instructional strategy.

The F ratios resulting from the models for each main effect were tested with the probability of falsely rejecting a hypothesis set at 5%. The procedures employed to test pairwise comparisons of means for significant F ratios were Student Newman Keul's and Duncan's simultaneous significant post-hoc tests.

Procedure

The administration of each of the five instructional strategies followed a common procedure consisiting of four phases.

Phase I

Distributed a set of instructional objectives to each subject of each treatment group, as well as the subjects in the control group.

Phase II

The subjects were given the package pretest for a maximum period of 30 minutes. The subjects' length of time for testing was recorded on his individual test by the monitor for analysis in this study.

Phase III

Subjects were given a set of instructional materials appropriate to their randomly assigned treatment group. The maximum length of instructional time was 50 minutes for each group. When a subject completed the materials he returned the materials to the monitor. The monitor recorded the amount of time for each subject.

Phase IV

Immediately following the treatment the subjects were given a post-test on the materials. The maximum length of time for the post-test was 30 minutes. The same procedure for recording the subjects' time for taking the post-test was used as was used in the pre-test setting.

Results

The means, standard deviations, and Ns for the five instructional strategies and the control group on the dependent variables total package time, post-test time, teaching time, pre-test score, post-test score and adjusted post-test score are reported in Table 1.

The dependent variables total package time, post-test time and teaching time were analyzed by an analysis of variance model (ANOVA). The results of these analyses indicated a significant difference existed between the groups; total package time, $F[.95]\ (4, 595) = 59.457$, $p < .05$; post-test time, $F[.95]\ (5, 719) = 20.68$, $p < .05$; teaching time, $F[.95]\ (4, 595) = 67.51$, $p < .05$. The pairwise comparisons were analyzed for each of these dependents using the Student Newman Keul's (SNK) post-hoc simultaneous significant test. The results of the SNK tests are condensed and reported in Table 2. The results indicate the text and question group was not significantly higher than the programmed instruction group in the dependent variables post-test time, teaching time or total package time. Thus, an instructional strategy has been identified that is as efficient as programmed instruction.

Evaluation of the effectiveness of the instructional strategies was measured in terms of performance.

The dependent variables pre-test and post-test performance scores were analyzed by an analysis of co-variance model (ANCOVA). The result of this analysis indicated a significant difference existed between the adjusted post-test performance scores reported in Table 1 with an $F[95]\ (5, 713) = 53, 59$, $p < .05$. A test of homogenity of regression was performed and no violation of the assumption was observed. The pairwise comparisons were analyzed using the Duncan post-hoc simultaneous significant test. The results are shown in Table 3 and indicate the test plus question group was not significantly different than the programmed instruction group. Thus, an instructional strategy has been identified that is as effective as programmed instruction.

The actual formation of the five instructional strategies consisted of incorporating the four components: content, questions, KR, and remediation into the instructional development strategies as described below.

1) Programmed Instruction: Subject was given a self-instructional set of materials that had been developed to move the subject through the material.

2) Text: Subject was given a manual that contains the content material with no programming of the content material.

3) Text/Questions: Subject was given a manual that contains the content material plus a set of questions which were at the end of the material. The directions preceding the questions instructed the learner to overtly respond to the questions.

4) Text/Questions/Knowledge of Results: Subject was given a manual and a set of questions as above. This instructional strategy was designed to include the correct responses for the questions in addition to text/questions.

5) Text/Questions/KR/Remediation: Subject was given a text containing the content material, questions, KR, and directions to portions of the text to review non-mastered content material.

6) Control: Subject was pretested and post-tested with no instructional treatment.

Instruments

The pre-test and post-test consisted of 27 matching items. The format was to provide each subject with a list of statements and a graphic representation of the low-voltage circuit tester. The graphic included the components of the low-voltage circuit tester pictured with numerics. The objective was to match the statement with the numeric in the graphic.

Discussion

The findings suggest an alternative instructional strategy has been developed that is as effective and efficient as programmed instruction. Programmed instruction results from a rigorous application of the ISD procedures and is also extremely time-consuming and therefore expensive. The instructional strategy which has been found to be as effective and efficient as programmed instruction uses the principles of mathemagenic behavior and auto-elucidation techniques. The application of mathemagenic activities and auto-elucidation techniques have cost effective implications for the conversion of instructional materials to self-instructional materials.

Conclusions

This research has examined alternative instructional development strategies and auto-elucidation techniques, e.g., effects of KCR, remediation, and post-text questions. This investigation has focused on the effectiveness of five types of instructional strategies.

The results support the application of auto-elucidation techniques in the form of post-text questioning. Text or written material followed by post-text questioning resulted in performance as effective as programmed instruction. Post-text questions are therefore regarded as aids to the learner which stimulate the learning environment.

Post-text questions supplementing but following the reading of written prose or text have been demonstrated as an effective mixture of instructional materials. For example, the treatments text, text plus questions plus feedback or text plus questions plus feedback plus remediation when uncontrolled was found to be less effective than P.I. or the use of text plus post-text questions.

In summary, the results support the following conclusions: 1) the use of post-text questions in the form of text and post-text questions was demonstrated to be the most effective alternative strategy; 2) the application of auto-elucidation components in the form of knowledge of results and remediation when uncontrolled was less effective than the use of post-text questions only.

IMPLICATIONS AND RECOMMENDATIONS

The primary purpose of this investigation was to identify the efficiency of five types of instructional development strategies.

This study investigated the effects of applying auto-elucidation techniques which were developed by Sidney Pressey. E.Z. Rothkopf and L. Frase further studied the process of incorporating questions into written prose. The implications are that auto-elucidation techniques in the form of post-text questions create a learning situation that is both meaningful to the learner and desirable to the instructor and the institution. The meaningfulness of the learning situation has been described in the literature by E.Z. Rothkopf and L. Frase as mathemagenic behavior. The use of post-text questions requiring overt response without any form of feedback produced the most desirable results both in terms of time and performance for the alternative instructional strategies.

This implication has major impact in the psychological area of feedback for curriculum developers, instructional technolgists and the field of psychology. Arguments against this implication would generally indicate that treatment groups receiving text + questions and feedback would presumably obtain higher scores. However, as demonstrated within this study, this was not the case. Therefore, instructional materials for which treatment groups do not have access to the correct answers but are permitted to review the instructional materials, are both effective and efficient.

Mathemagenic behavior as described in the literature and discussed in this study can be present in the absence of informational feedback.

The instructional strategies investigated have incorporated the principles of systems analysis, informational feedback and auto-elucidation processes.

Instructional development models used in this study yielded instructional materials that were efficient in terms of reducing teaching time and effective in terms of performance. Therefore, the conversion of existing instructional materials into self-instructional modules using mathemagenic principles can have cost-benefit advantages for educational, industrial and Armed Service institutions.

The implications for cost-effective instructional materials development are as follows:

1) Material may be developed that yield performance scores higher than existing ISD developed materials while requiring less teaching time;

2) The cost associated with development of the materials may be significantly less.

The next implication concerns the Armed Services. The Armed Services have two plans under which all training operates. The two plans are peacetime and wartime. Without discussing either of the plans in detail, the fact remains that during wartime the Armed Services must reduce the training time of soldiers in order to meet the demands. Therefore, considerations must be made regarding the instructional materials. Results of this study indicate that the difference in the means for text (2) and question (3) groups total teaching time was 6.76 minutes which represents one-third less time. The decrease in performance was 3.37 test items or 12% of the total possible score. Without discussion about the future effectiveness of the soldier, the point to be considered is that alternative instructional strategies may cut training time by 1/3, therefore better meeting the wartime needs of our country.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF TEACHING TIME, POST-TEST TIME, TOTAL PACKAGE TIME

PRE-TEST SCORE, POST-TEST SCORE, AND ADJUSTED POST-TEST SCORES

BY INSTRUCTIONAL STRATEGY

GROUPS	TEACHING TIME			POST-TEST TIME		TOTAL PACK. TIME		PRE-TEST SCORE		POST-TEST SCORE		ADJ.
	n	M	SD	M	SD	M	SD	M	SD	M	SD	M
PI	120	37.40	9.95	14.00	7.10	51.40	9.9[illegible]	3.14	3.54	12.00	6.50	12.93
T	120	22.50	8.40	11.55	5.25	34.04	10.17	1.01	2.63	10.53	5.76	11.05
TQ	120	38.05	7.96	12.75	7.00	50.80	10.17	2.52	2.00	14.21	6.00	14.42
TQF	120	29.42	7.55	11.05	5.42	40.46	10.37	3.13	3.51	12.42	5.10	11.96
TQFR	120	33.04	8.24	10.67	6.96	43.72	11.22	2.70	2.61	10.59	6.70	10.56
CTRL	120	*	*	6.73	4.70	*	*	3.00	2.67	4.12	4.02	3.70

* = TREATMENT NOT ADMINISTERED TO CONTROL GROUP

TABLE 2

PAIRWISE COMPARISON OF MEAN DIFFERENCES WITH
FIVE INSTRUCTIONAL TREATMENTS ON TOTAL PACKAGE TIME

q(r,595)		2.80	3.36	3.69	3.92
Ordered Group Mean	2	4	5	3	1
(2) Test	-	*	*	*	*
(4) T+Q+F		-	*	*	*
(5) T+Q+F+R			-	*	*
(3) Test+Quest				-	ns
(1) PI					-

* = $p < .05$, q(r,595) $\sqrt{MSe/n}$ = q(r,595)([illegible])

PAIRWISE COMPARISON OF MEAN DIFFERENCES WITH
FIVE INSTRUCTIONAL TREATMENTS ON POST-TEST TIME
AND THE CONTROL GROUP

q(r,595)		2.80	3.36	3.69	3.92	4.10
Ordered Group Mean	6	5	4	2	3	1
(6) Control	-	*	*	*	*	*
(5) T+Q+F+R		-	ns	ns	*	*
(4) T+Q+F			-	ns	ns	*
(2) Test)				-	ns	*
(3) T+Quest					-	ns
(1) PI						-

* = $p <$ .05, q(r,714) $\sqrt{MSe/n}$ = q(r,714)([illegible])

PAIRWISE COMPARISON OF MEAN DIFFERENCES
WITH FIVE INSTRUCTIONAL TREATMENTS ON TEACHING TIME

q(r,595)		2.80	3.36	3.69	3.92
Ordered Group Mean	2	4	5	1	3
(2) Test	-	*	*	*	*
(4) T+Q+F		-	*	*	*
(5) T+Q+F+R			-	*	*
(1) PI				-	ns
(3) Test+Quest					-

* = $p < .05$, q(r,595) $\sqrt{MSe/n}$ = q(r,595)(.774)

TABLE 3

PAIRWISE COMPARISON OF MEAN DIFFERENCES OF THE SIX GROUPS ON ADJUSTED POST-TEST SCORE AND THE CONTROL GROUP

q(r,713)		2.80	2.95	3.05	3.12	3.22
Ordered Group Mean	6	5	2	4	1	3
(6) Control	-	*	*	*	*	*
(5) T+Q+F+R		-	ns	ns	*	*
(2) Text			-	ns	ns	*
(4) T+Q+F				-	ns	*
(1) PI					-	ns
(3) T+Quest						-

* = p .05, q(r,713 [**] = q(r,713)(.71238)

**NOTE: The adjusted MSe is an estimate of the average standard error (Dayton, 1970, pp. 322).

$$q = \sqrt{\frac{2MSerror^*}{n}\left[1 + \frac{SS\ treat}{(p-1)\ SSx\ error}\right]}$$

Course Quality Measurement of CAI

Edward F. Magdarz

INTRODUCTION

The IBM Field Engineering Division uses a CAI/CMI system with approximately 400 terminals in branch offices across the country to accomplish technical training of IBM's customer service personnel. The system currently administers 374 different courses and averaged 28,000 student hours per month in 1976. Many new courses are developed each year to keep pace with advances in product and service technology.

This CAI/CMI system is called the Field Instruction System - Version II (FIS II). It uses a modification of an IBM programmed product, the Interactive Instruction System (IIS) as the teaching vehicle. Another IBM programmed product, Interactive Query and Report Processor (IQRP) is utilized for the course quality analysis part of FIS II. The application of these two programmed products (IIS and IQRP) is how FIS II accomplishes "course quality measurement of CAI".

FIS II utilizes an internal company teleprocessing network to connect more than 160 branch offices to a central computer in New York.

Course quality is a major concern in this system because of the large number of new courses released each year and because of the decentralized structure of the course development organization. Fifty-one new courses, consisting of 140 units that average five hours each, were released in 1976. They were developed by twelve geographically separate departments.

The purpose of this publication is to describe a course quality measurement system. It is necessary to establish a proper context for this description, therefore the first few pages will be used to establish an overall picture of FIS II and the functions which support the course quality measurement part of the system.

HISTORY

The Field Engineering Division became involved in CAI in the early 1960's when "book type" programmed instruction material was adapted to CAI typewriter terminals. This was an experimental era. Several CAI systems were tried and evaluated and research was conducted to identify the most effective types of CAI materials and learning methods/models. In 1968 a sophisticated full scale teleprocessing CAI system was implemented.

Backed by seven years experience, a redesigned CAI system was implemented in 1975. The IBM Interactive Instructional System was selected as the base for the redesigned system. IBM 3270 Information Display System cathode ray tube terminals replaced the typewriter terminals and a new improved course model was developed. The new model incorporates current innovations in training psychology and the experience gained from 4,000,000 student hours executed on the previous system. This new CAI system was named Field Instruction System - Version II (FIS II).

FIS II course structure follows a typical CAI model imploying multi-level components, pre/post testing, remedial and unit (criteria) testing. A brief description of each component follows (Figure 1):

Course: Average course length is 10 hours, ranging from 1 to 80 hours. Satisfactory completion of a course certifies that the individual can perform certain job tasks when servicing customer equipment. Some FIS II courses are prerequisite to Education Center laboratory courses where actual "hands on" experience is gained.

Unit: This is a logical sub-component of the course's subject. Average length is 5 hours. During Unit Evaluation at the end of each unit the students opinion is solicited via a Student Opinion Questionnaire an a Unit Test is administered. Course quality measurement data is stored "by unit" resulting in the capability to analyze the quality of each unit.

Session: A session is a logical grouping of objectives. Each objective has a related:

- Test item
- Assignment (teaching material)
- Remedial material

Correctly answering a Pretest Item branches the student past the related Assignment and past any further testing (Post Test and Unit Test).

For each Assignment presented to the student a Post Test Item is administered. Remedial material is administered for Post Test Items not answered correctly.

Assignment: The Assignment teaches a single objective and is composed of Activities, each relating to a separate student learning task such as:

- Reading text (on the screen or referenced in another medium)

- Answering study questions
- Working on application type exercises
- Solving a problem

The Assignment and Activity levels are where the "teaching" occurs.

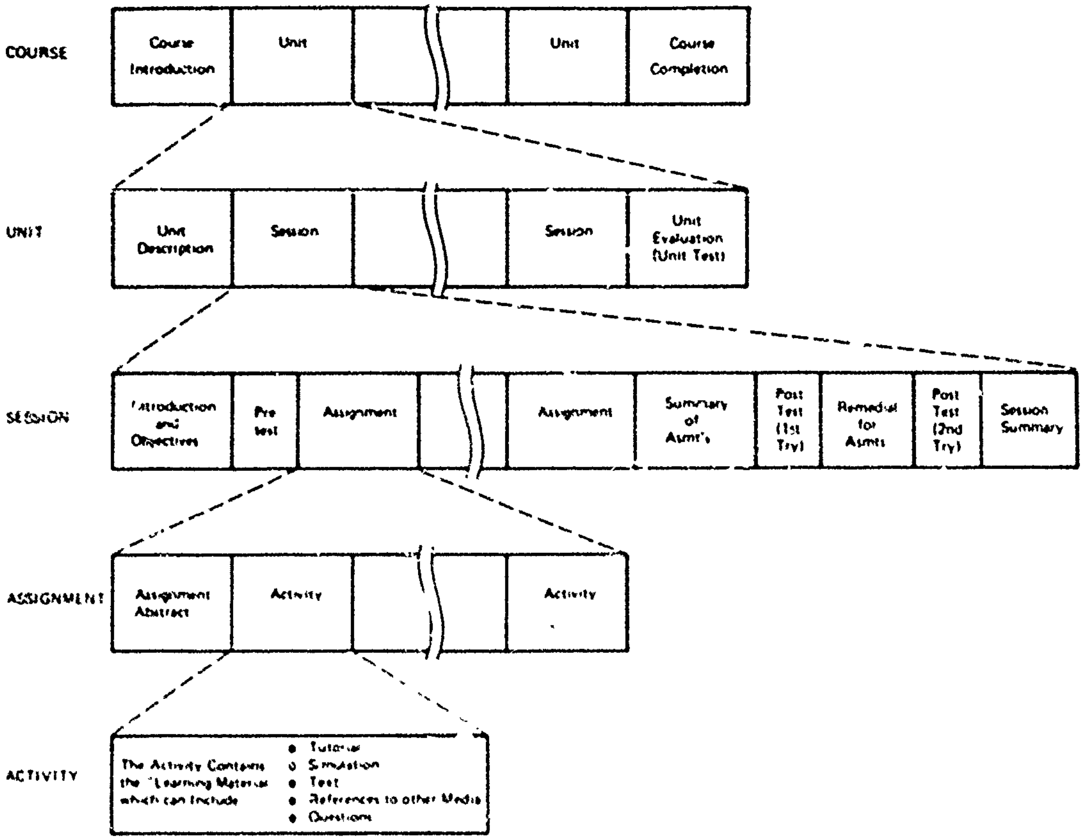

COURSE
Course Introduction
Unit
Unit
Course Completion
UNIT
Unit Description
Session
Session
Unit Evaluation (Unit Test)
SESSION
Introduction and Objectives
Pre test
Assignment
Assignment
Summary of Asmt's
Post Test (1st Try)
Remedial for Asmts
Post Test (2nd Try)
Session Summary
ASSIGNMENT
Assignment Abstract
Activity
Activity
ACTIVITY
The Activity Contains the "Learning Material which can Include
Tutorial
Simulation
Test
References to other Media
Questions

FIGURE 1

MEASUREMENT COMPONENTS

Common Test Items are used for the various tests in each unit. Typically, two alternate-equivalent items are developed for each objective in the unit. One of the items is randomly selected and administered, as appropriate, in the Session Pretest, Session Post Test and Unit Test. Using common items reduces the authors test development effort and helps ensure that each test is valid to it's objectives.

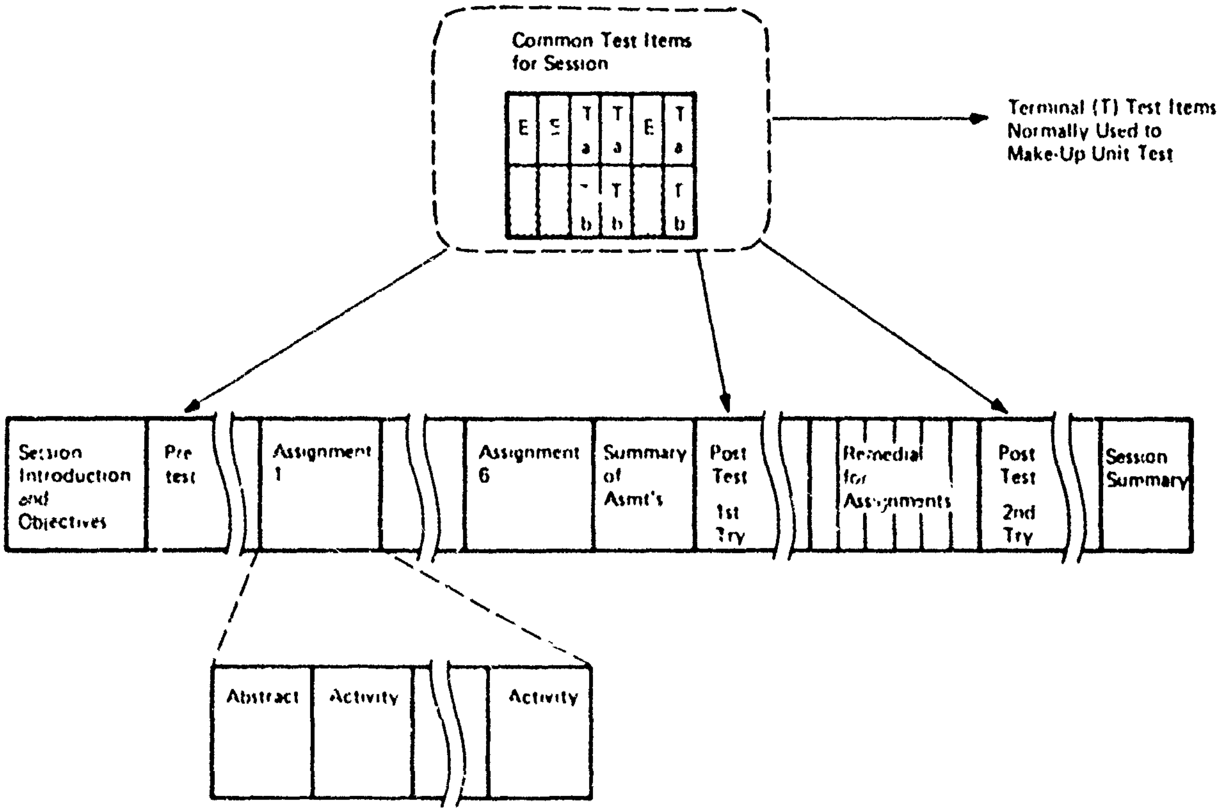
Alternate Test Items are used because the student may be tested several times on the same objective. Administering an alternate-equivalent item during the re-testing of an objective accomplishes two things:

- Reduces student feelings of "continually seeing the same questions".

- Reduces the possibility of the student correctly answering a re-test item from information temporarily memorized from previous testing.

Test Item = Objective: Each test item measures a single objective. A test item can consist of any number and type of questions, exercises, problems, etc. This ensures that an item can comprehensively measure it's objective.

Two Level Testing reduces the total number of test items administered to the student while maintaining comprehensive testing capability. For example, a terminal objective has two enabling objectives. When learning the three objectives the student is first given teaching material for the two enabling objectives. Then, the teaching material for the terminal objective is presented. However, during testing the student is given the terminal test item first and if it is answered correctly the two enabling test items are skipped.



The example shows the content of a typical session. This session is based on six objectives, including both Terminal (T) and Enabling (E) objectives and related test items. The terminal test items have "a" and "b" alternate versions.

FIGURE 2

Student Opinion Questionnaire solicites the student's opinion of the course, the system (FIS II) and of the local branch office study environment. It is administered at the end of each unit and uses a three level branching logic which is designed to (Figure 3):

- Continue to solicit opinion details if the student's responses are significantly positive or negative.

- Stop soliciting opinion details when the student stops responding.

At the first level every student is asked for an overall opinion of the unit and must respond to a six choice range:

- Very Good

- Good

- Average

- Poor

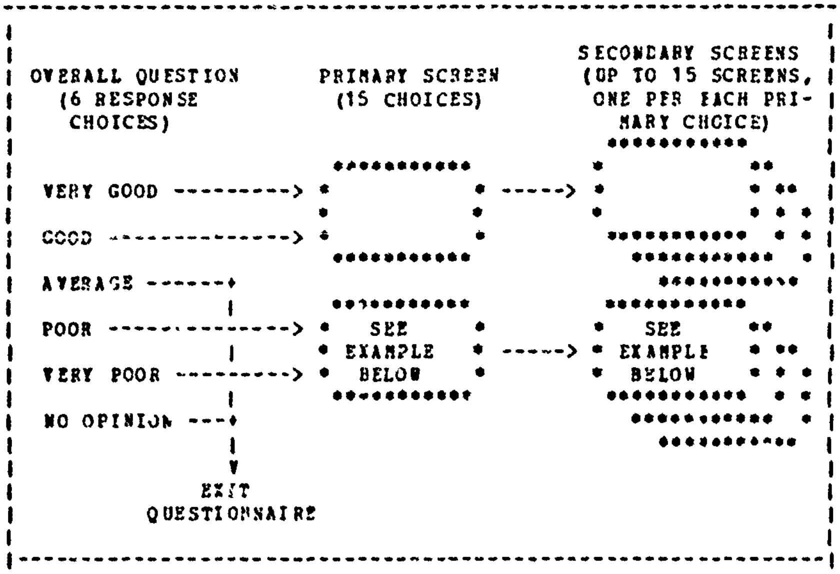
- Very Poor

- No Opinion

A response of Average or No Opinion ends the questionnaire. A response of Very Good, Good, Poor, or Very Poor takes the student to level two of the questionnaire where a single screen containing fifteen selectable items is displayed. If none of the items are selected the questionnaire ends. However, for every item that is selected a level three screen is displayed containing additional detailed items.

This branching type questionnaire minimizes the amount of student time spent in the questionnaire and reduces the collection and analysis of "null" responses. The maximum number of of responses the student could give, including all items in all three levels is seventy-five. The minimum number is one, a null response to the level one question.

SAT/UNSAT Grading: The unit test, and all teaching material, can be re-taken as many times as necessary to satisfactorily complete the test. A satisfactory Course Completion Certificate is given by the system when all unit tests have been satisfactorily completed.



EXAMPLE OF PRIMARY SCREEN CHOICES
You indicated, "the unit is POOR/VERY POOR", what are your reasons?
1. COVERAGE of the subject is wrong (too little/too much)
2. MATERIAL presented is DIFFICULT to UNDERSTAND
3. HUMOR and use of INFORMAL language is poor
. . . . .
. . . . .
15. SUBJECT being taught is NOT INTERESTING
x. REASON is NOT SHOWN above

EXAMPLE OF CHOICES ON A SECONDARY SCREEN
You indicated, "COVERAGE of the subject is wrong", what are your reasons?
1. TOO LITTLE is covered - I don't feel confident
2. TOO MUCH is covered - unnecessary material is presented
x. REASON is NOT SHOWN above

FIGURE 3

MEASUREMENT SYSTEM

<u>Concept</u>: When a new course is released, how can you tell if it's a "good" course or if it's a "bad" course? FIS II uses the term Course Quality to describe how good or bad a course is. Course quality consists of three components (Figure 4):

- Effectiveness (ability of student to perform as stated in objectives)
- Efficiency (ability of student to learn in minimal time)
- Acceptance (student opinion of the training experience)

There is a considerable amount of inter-relationship between the three components. For example, a student who "likes" the training experience may learn "faster". The components in this system constitute a hierarchy of importance and if the system was limited to measuring only one component it would be effectiveness.

<u>Users</u>: The system is designed for these users:

- Author (staff personnel who develops the CAI course or course unit)
- Point of Control (local management and senior staff personnel who have on-going responsibility for a group of courses)
- Headquarters (upper management - responsible for all courses)
- Research

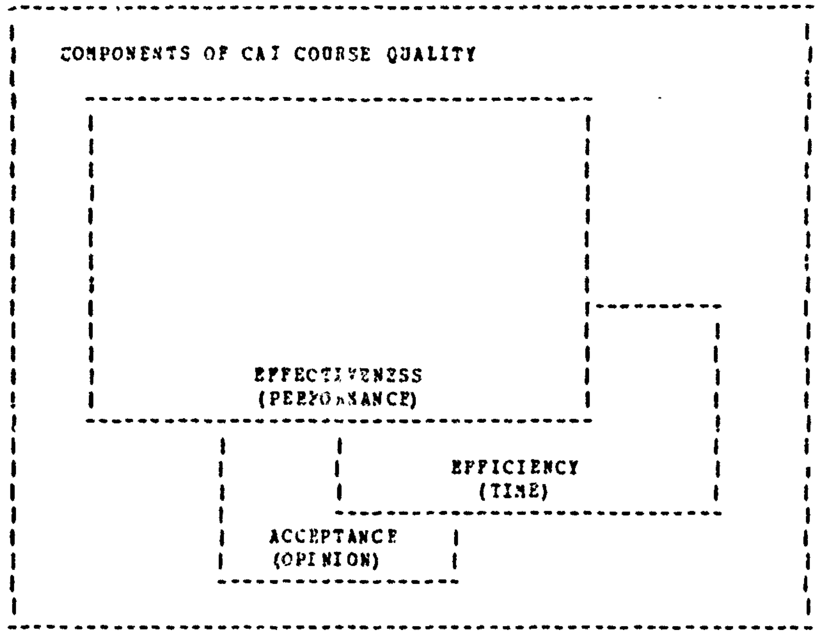

COMPONENTS OF CAI COURSE QUALITY
EFFECTIVENESS
(PERFORMANCE)
EFFICIENCY
(TIME)
ACCEPTANCE
(OPINION)

FIGURE 4

__Other Features__: The system is a "demand only" system. Each report must be requested. Reports do not have a scheduled printing and distribution cycle. This keeps paperwork to a minimum and helps ensure the course quality information ends up only in the hands of those who want it, when they want it.

All reports are displayable/printable on-line at any of the systems CAI terminals. Special requests for analysis of unique combinations of data can be formulated and the report can be obtained on-line in minutes. This is especially useful when seeking the cause of a "low" quality problem where repeated analysis is required.

__Uses__: The system is used many ways by each of the users. Primary uses include:

- Reinforce author.

- Identify low quality training requiring revision.

- Identify quality training methods and techniques for propagation.

- Allow management to direct revision resources to the lowest quality courses.

During design of the FIS II measurement system existing programs were examined, for suitability, in an attempt to eliminate the development of a new program. The Field Engineering Management Information System (FE/MIS), an internal IBM program met the design specifications and was selected. This program is now available outside IBM as a programmed product called Interactive Query and Report Processor (IQRP).

The IQRP features that are most applicable to the FIS II Course Quality measurement system include:

- Record selection flexibility
- Calculation capability
- Sequencing/subtotaling capability
- Output format flexibility
- "Macro" capability for "packaging" often used reports

Record Content: The record obtained from each student during the completion of each course unit is 290 characters long and includes the following course quality data:

IDENTIFICATION DATA

- Course and unit number
- Employee and branch office number
- Date of unit completion
- Point of Control location responsible for course

TEST DATA

- Number of test tries
- Result of last test try (satisfactory or unsatisfactory)

- Item results (students response to each of a maximum of 81 Unit Test items)

TIME DATA

- Student's "on terminal" study time

- Student's "off terminal" study time (as reported by student)

- Student's "on terminal" Unit Test time

STUDENT OPINION QUESTIONNAIRE DATA

- Student's overall rating of the unit (Very good, good, average, poor, very poor, no opinion)

- Student's responses to detailed opinion items (maximum of 74 items)

Rolling History: Only the 50 most current records (students) are stored for each course unit. When a record for a new student completion is put on the IQRP file, the oldest (51 st) record for that unit is purged. This results in a reasonable storage "size" requirement and still provides current usable data for all users. Currently the FIS II IQRP file contains over 9,000 records and includes 160 units from 70 courses. The file continues to grow as new courses are released.

Questions - Inquires: IQRP is used by formulating questions about the data and analysis wanted (Figure 5). These questions are entered at a terminal in an abbreviated language called the IQRP inquiry (examples are shown later in the RESULTS section).

EXAMPLE OF QUESTIONS USED IN FORMULATING ICFP INQUIRIES

What is the student acceptance of Unit #2 of Course #700339?

What are reasons that cause some students to indicate that Unit #2 of Course #70339 is not acceptable?

What is the average student acceptance for Course #70339 (all units)?

What is the average student acceptance for all units in all courses developed by department NYC?

Do students who rate units as "very acceptable" complete the units faster or slower than students who rate units as "not acceptable"?

NOTE: The preceding questions all involved the "acceptance" component of Course Quality. Similar questions can also be asked concerning the "effectiveness" and "efficiency" components.

FIGURE 5

METHODS USED

Four methods are used to analyze course quality (Figure 6):

- IQRP Summary
- IQRP Detail
- Response Recording by Unit
- Response Recording by Student

These methods form a hierarchy of "level of information detail", with Response Recording being the most detailed. The methods using IQRP yield generalizable results, such as percentages, that allow comparison of unit-to-unit, course-to-course, and Point of Control-to-Point of Control. The remainder of this publication will concentrate on the IQRP methods.

The Response Recording analysis methods are not generalizable. They yield results, such as specific student answers, that cannot be compared to other units, courses, or Points of Control. These analysis methods are a built-in feature of IBM's Interactive Instruction System where they are called Course Activity Summary report and Response Recording report.

The ability to compare the quality of units, courses, and Points of Control is very important to the Field Engineering Division because of the decentralized structure of the departments that develop the courses.

```
-------------------------------------------------------------------------------
| FIS2 TIME/UNIT COURSE = 70339 SUBTL/LH UNIT T/O.
| PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
| FILE UPDATED ON 08/15/77
| COURSE C AUTH PLAN                                        UNIT/AVG    PCT/AVG
|        U LCC  TIME
|
| 70339  1 NYC   9.2                                          8.1         87
| RECORDS   50
| 70339  2 NYC   2.1                                          2.2         106
| RECORDS   50
|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/|___
```

IQRP SUMMARY EXAMPLE: Each line shows Unit "summary" data (50 students) for effectiveness, efficiency, or acceptance.

```
-------------------------------------------------------------------------------
| FIS2 TIME COURSE = 70339 UNIT = 1 LIMIT 55.                                 |
| PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***                           |
| FILE UPDATED ON 08/15/77                                                    |
| COURSE C AUTH  SERIAL B/O  COMP DTD  C L PLAN  TIME  TIME  FST  LST |
|        U LOC                         P _ TIME  ON    OFF   TST  TST |
|                                        #                   TRY  TRY |
| 70339  1 NYC  *70270 836  08/15/77  S 1  9.2  10.0   1.5   .5   .5 |
| 70339  1 NYC  *75235 836  08/15/77  S 1  9.2   6.4    .2   .2   .2 |
| 70339  1 NYC  370076 422  08/13/77  S 2  9.2   8.6    .4   .6   .2 |
|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/
```

IQRP DETAIL EXAMPLE: Each line shows Unit "detail" data (1 student) for effectiveness, efficiency, or acceptance.

```
-----------------------------------------------------------------
| RESPONSE RECORDING BY UNIT                                     |
| COURSE 70343          UNIT 7                                   |
| LABEL STUDENTS  CA   WA                                        |
|                                                                |
| 61?401   86      0    0    (REMEDIAL - AFTER 1ST POST TEST TRY) |
| 61C411    8      1    7    (PRETEST ITEM)                      |
| 61E4??  1?5      0    ?    (ABSTRACT - START OF TEACH MATERIAL) |
| 61M411   5?     56   26    (POST TEST ITEM - PART I)           |
|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/|___/
```

RESPONSE RECORDING BY UNIT EXAMPLE: Each line shows accumulated student responses occurring at a single display screen.

```
---------------------------------------------
| Report too detailed to illustrate here.
|___/|___/|___/|___/|___/|___/|___/|___/
```

RESPONSE RECORDING BY STUDENT: Each report shows every response made by a student.

FIGURE 6

RESULTS

The report examples used in this section are condensed and simplified. However they do represent actual or typical results of the FIS II Course Quality Measurement system.

REPORT EXAMPLES

IQRP Summary Example: The top line of each report is the IQRP language inquiry that was keyed in at a terminal and produced the on-line report. The inquiry for the first report shown translates to (Figure 7):

- Asks IQRP to look at FIS II data only (FIS2).
- Requests a preformatted report called TEST/RESULT.
- Asks only for courses within limits (WL) of Course #70335 and Course #70350. Only two of the courses are shown.
- Asks for subtotals by unit and course (SUBTL) and to arrange courses and units in sequence from low-to-high (LH).
- Asks for totals only (T/O), rather than a report line for each of the 50 students on file.

Now lets examine the report data. First the column headings and Line 1 of the data:

- COURSE = The first data line is for Course #70339.
- CU = Data on this first line is for course Unit #1.
- AUTH LOC = Department responsible for this unit is NYC.
- PASS FAIL = Criteria in %, for passing Unit Test, as set by author.
- SAT = Number of students who completed the Unit Test satisfactorily (50 of 50).

FIS2 TEST/RESULT COURSE WL 70335,70350 SUBTL/LH COURSE,UNIT T/O.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LOC	PASS FAIL	SAT	UNSAT	TRY/1	TRY/2	TRY/3+	SAT1&2/PCT	SAT/ADJ
70339	1	NYC	70	50		50			100	70
RECORDS		50								
70339	2	NYC	50	50		47	2	1	98	48
RECORDS		50								
70339	3	NYC	100	50		50			100	100
									99	73
RECORDS		50;	150							
70343	1	SPC	70	44	1	42	2		98	68
RECORDS		45								
70343	2	SPC	70	45		30	9	6	87	57
RECORDS		45								
70343	3	SPC	70	45		23	9	12	71	41
									85	55
RECORDS		45;	135							

FIS2 TIME/UNIT COURSE WL 70335,70350 SUBTL/LH COURSE,UNIT T/O.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LOC	PLAN TIME	UNIT/AVG	PCT/AVG
70339	1	NYC	9.2	8.1	87
RECORDS		50			
70339	2	NYC	2.1	2.2	106
RECORDS		50			
70339	3	NYC	10.0	14.1	141
					111
RECORDS		50;	150		
70343	1	SPC	16.0	16.8	105
RECORDS		45			
70343	2	SPC	11.0	8.9	80
RECORDS		45			
70343	3	SPC	13.0	12.0	92
					92
RECORDS		45;	135		

FIGURE 7

- UNSAT = Number of students who have not yet completed Unit Test satisfactorily (0 of 50).

- TRY/1 = Number of students who have attempted the Unit Test only once (50 of 50).

- TRY/2 = Number of students who have attempted the Unit Test twice.

- TRY/3+ = Number of students who required three or more attempts.

- SAT1&2/PCT = % of students who completed Unit Test satisfactorily in first two tries (100%).

- SAT/ADJ = Adjusted SAT1&2/PCT column. Adjustment is made by lowering the SAT1&2/PCT figure proportional to the PASS FAIL column to reflect the affect of a test with a lowered criteria (70). The SAT/ADJ figure is the number that FIS II uses as the Course Quality - Effectiveness indicator.

Now lets examine the overall report result for Course #70339. Unit #1 has an effectiveness indicator of 70. Unit #2 is 48 and Unit #3 is 100. The indicator of average effectiveness for the entire course is 73. The cause of the low effectiveness indicator for Unit #2 (48) should be investigated and improvements made.

Notice that the RECORD COUNT for each unit of Course #70339 is 50. This means the course is not a "brand new" course and the "rolling 50" storage function is effect. The data shown on this report is for the most recent 50 students. In contrast, look at the RECORD COUNT for Course #70343. It is a new course and has less than 50 students.

The second report shown is the Course Quality measurement of Efficiency and is called TIME/UNIT. The PCT/AVG column shows, the average unit completion time of the students on file (column UNIT/AVG), as a percentage of the planned unit time (column PLAN TIME).

For Course #70339, Unit 1, the actual average unit completion time is 8.1 hours which is 87% of the planned unit time of 9.2. This unit appears efficient because the students are completing faster than planned.

FIS2 TEST/RESULT COURSE WL 70335,70350 SUBTL/LH COURSE,UNIT T/O.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LCC	PASS FAIL	SAT	UNSAT	TRY/1	TRY/2	TRY/3+	SAT1&2/PCT	SAT/ADJ
70339	1	NYC	70	50		50			100	70
RECORDS		50								
70339	2	NYC	50	50		47	2	1	98	48
RECORDS		50								
70339	3	NYC	100	50		50			100	100
									99	73
RECORDS		50;	150							
70343	1	SFC	70	44	1	42	2		98	68
RECORDS		45								
70343	2	SFC	70	45		30	9	6	87	57
RECORDS		45								
70343	3	SFC	70	45		23	9	12	71	41
									85	55
RECORDS		45;	135							

FIS2 TIME/UNIT COURSE WL 70335,70350 SUBTL/LH COURSE,UNIT T/O.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LCC	PLAN TIME	UNIT/AVG	PCT/AVG
70339	1	NYC	9.2	8.1	87
RECORDS		50			
70339	2	NYC	2.1	2.2	106
RECORDS		50			
70339	3	NYC	10.0	14.1	141
					111
RECORDS		50;	150		
70343	1	SFC	16.0	16.8	105
RECORDS		45			
70343	2	SFC	11.0	8.9	80
RECORDS		45			
70343	3	SFC	13.0	12.0	92
					92
RECORDS		45;	135		

FIGURE 7

The third report shows Course Quality - Acceptance in the form of responses to the Student Opinion Questionnaire (SOQ) (Figure 8 - Top). The RATING column shows the arithmetic mean of the responses to an overall opinion question, on a scale of five ranging from Very Good (index of 5) to Very Poor (index of 1) plus a null response of No Opinion. If all students rated a unit as Average the RATING would be 3.00.

IQRP Detail Example: The preceeding TIME/UNIT report showed "summary" data. It would normally be used to get a general idea of course efficiency. The data that makes up the summary report can be shown on a "detail" report (Figure 8 - Bottom). The example shows TIME detail that would normally be used by the course author when investigating a problem or by research personnel. Each of the 50 students on file is shown as an individual report line. Shown for each student is; study time on terminal, study time off terminal, and time for the first and last unit test tries.

Similar "detail" data reports are available showing results of each test item (effectiveness) and responses to the 74 detailed opinion questions (acceptance).

FIS2 SOQ/RATE COURSE WL 70335,70350 SUBTL/LH COURSE,UNIT T/O.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LOC	A=5 (VG)	B=4 (G)	C=3 (A)	D=2 (P)	E=1 (VP)	X=3 (NC)	RATING	DE/PCT
70339	1	NYC	4	21	14	5		6	3.48	10
RECORDS		50								
70339	2	NYC	3	19	16	5	2	5	3.32	14
RECORDS		50								
70339	3	NYC	3	19	16	5		7	3.40	10
			10	59	46	15	2	18	3.38	11
RECORDS		50; 150								
70343	1	SPC	5	20	15	1		4	3.64	2
RECORDS		45								
70343	2	SPC		3	22	15	2	3	2.64	33
RECORDS		45								
70343	3	SPC	3	14	20	3		5	3.37	7
			8	37	57	19	2	12	3.22	16
RECORDS		45; 135								

FIS2 TIME COURSE = 70339 UNIT = 1 LIMIT 55.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LOC	SERIAL	B/O	COMP DTD	C Y	L T #	PLAN TIME	TIME ON	TIME OFF	FST TST TRY	LST TST TRY
70339	1	NYC	*70270	836	08/15/77	S	1	9.2	10.0	1.5	.5	.5
70339	1	NYC	*75235	836	08/15/77	S	1	9.2	6.4	.2	.2	.2
70339	1	NYC	370076	4?2	08/13/77	S	2	9.2	8.8	.4	.6	.2
70339	1	NYC	821466	095	06/28/77	U	2	9.2	11.3	.7	.3	.4
70339	1	NYC	421590	155	06/27/77	S	1	9.2	7.1	.1	.2	.2
RECORDS		50										

FIGURE 8

EXAMPLE OF AUTHOR USE

A new course has been developed and the author is interested in it's resultant quality.

Step 1: The author obtains the IQRP summary reports for TEST/RESULT (effectiveness), TIME/UNIT (efficiency), and SOQ/RATE (acceptance) via a terminal. In this example only the SOQ/RATE report will be used. It shows low acceptance for Unit #2 (2.64). The report indicates that Units #1 and #3 are much more acceptable than Unit #2 (Figure 9).

Step 2: In order to investigate the low acceptance of Unit #2, the author obtains an IQRP detail report of student responses to the detail questions on the Student Opinion Questionnaire. Of the fifteen Primary reasons why students might not like Unit #2, Reason #4 has the most responses. The detail report shows:

- Of the 17 students who said the unit was Poor or Very Poor, 16 indicated Reason #4 ("...study activities are poor...") as a cause. They did not indicate any of the other fourteen Primary reasons, such as "poor tests" or "poor directions" as major causes of their low acceptance.

- The 16 students who said "study activities are poor" were administered a screen with the following Secondary reasons:

  - S22 Displayed study material not helpful in learning.

  - S23 Referenced study material not helpful in learning.

  - S24 Too many questions - problems - exercises.

  - S25 Too few questions - Problems - exercises.

  The report shows that most (14) of the dissatisfied students picked Secondary Reason #S23 as the cause. The author now suspects that there is a problem with the "referenced study material" in Unit #2.

- Unit #2 is several hours long, has several objectives, and has many different manuals that are referenced for study activity. Which manuals are the students complaining about?

FIS2 SOQ/RATE COURSE = 70343 SUBTL/LH UNIT T/O.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LOC	A=5 (VG)	B=4 (G)	C=3 (A)	D=2 (P)	E=1 (VP)	X=3 (NO)	RATING	DE/PCT
70343	1	SFC	5	20	15	1		4	3.64	2
RECORDS	45									
70343	2	SFC		3	22	15	2	3	2.64	33
RECORDS	45									
70343	3	SFC	3	14	20	3		5	3.37	7
			8	37	57	19	2	12	3.22	16
RECORDS	45;	135								

FIS2 PPS14 COURSE = 70343 UNIT = 2 T/O.
PG.1 08/18/77 *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77

COURSE	C U	AUTH LOC	PO4	S22	S23	S24	S25
70343	2	SFC	16	4	14		2
RECORDS	17						

RESPONSE RECORDING BY UNIT
COURSE 70343 UNIT 2

LABEL	STUDENTS	CA	WA	
61F401	86	0	0	(REMEDIAL - AFTER 1ST POST TEST TRY)
61C411				
61C412				
61E411				
61M411	52	56	28	(POST TEST ITEM - PART 1)
61M412	79	31	52	(POST TEST ITEM - PART 2)
61C421				
61C422				
61E421				
61E422				
61E423				
61E424	120	112	7	(END OF TEACH MATERIAL)
61M421	65	71	23	(ALTERNATE POST TEST ITEM - PART 1)
61M422	89	32	59	(ALTERNATE POST TEST ITEM - PART 2)

FIGURE 7

Step 3: The author obtains a Response Recording by Unit report via a terminal. This report shows accumulated student responses to EACH SCREEN, for EACH OBJECTIVE in Unit #2 (Figure 9 - Bottom). The data for Objectives #1 thru #3 was examined and looks OK. However, the data for Objective #4 (Shown in the example report) identifies a problem. The column headings CA and WA stand for "Correct Answer" and "Wrong Answer". This objective has fourteen display screens. Notice that at Screen (LABEL) #61m411 and #61m422 most students are responding with wrong answers rather than correct answers (61m412: CA's = 31 and WA's = 52). Further investigation on this objective solved the low acceptance problem. The manual reference for this objective was incorrect. The students were told to study one subject but were tested on another.

Summary: The author used reports at three levels of detail to identify a "low student acceptance" problem and pinpoint it's cause. The reports were available immediately, on-line, via a CAI terminal.

HEADQUARTERS TRACKING

Headquarters began using the Course Quality Measurement system to track "established" units in January 1977. Units that have had more than fifty student completions are considered "established". Here are the results of the first two quarters (1Q77 & 2Q77) of tracking:

- Number of courses currently being tracked = 33.
- Number of units currently being tracked = 104.
- Effectiveness:
  - TEST/RESULT criteria currently set at 75%.
  - Units meeting criteria = 62 of 104.
  - Best unit = 100%. Worst unit = 22%.
  - % of units in criteria 1st quarter = 54%.
  - % of units in criteria 2nd quarter = 60%.
- Efficiency:
  - TIME/UNIT criteria currently set at 75% to 110%.
  - Units meeting criteria = 51 Of 104.
  - Best unit = 100%. Worst unit = 196%.
  - % of units in criteria 1st quarter = 43%.
  - % of units in criteria 2nd quarter = 49%.
- Acceptance:
  - SOQ/RATE criteria currently set at 3.25.
  - Units meeting criteria = 44 of 104.
  - Best unit = 4.22. Worst unit = 2.42.
  - % of units in criteria 1st quarter = 40%.
  - % of units in criteria 2nd quarter = 42%.

EXAMPLE RESEARCH

The Course Quality measurement system provides timely answers to research type questions. Here are a few of the many questions answered thus far.

PLANNED UNIT TIME VS. ACCEPTANCE

FIS II has a course development guideline that recommends that the author design each unit to require eight hours, or less, learning time. Question: What is the relationship of student acceptance of units eight hours or less, and of units over eight hours? The two IQRP reports show that student acceptance appears better on units that are within the eight hour guideline (Figure 10 - Top):

- Planned unit time (P/UT) not greater than (NG) 8.0 hours results in acceptance (SOQ RATING) of 3.31.
- Planned unit time (P/UT) greater than (GT) 8.0 hours results in acceptance (SOQ RATING) of 3.21.

SLOW/FAST STUDENTS VS. ACCEPTANCE

Question: What is the relationship of student acceptance to how fast/slow the student learns? As the three reports show, it appears that students who complete quickly have a more favorable of the unit (Figure 10 - Bottom):

- Student completion time less than (LT) 75% of planned time (PCT/AVG) results in acceptance (SOQ RATING) of 3.40.
- Student completion time within limits (WI) of 75% and 125% of planned time (PCT/AVG) results in acceptance (SOQ RATING) of 3.26.
- Student completion time greater than (GT) 125% of planned time (PCT/AVG) results in acceptance (SOQ RATING) of 3.08.

```
FIS2 SOQ/RATE P/UT NG 8.0 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH         A=5   B=4   C=3   D=2   E=1   X=3   RATING    DE/PCT
       U LOC          (VG)  (G)   (A)   (P)   (VP)  (NO)

                      424  1996  2711   624   178  1263   3.31        11
RECORDS  7209
```

```
FIS2 SOQ/RATE P/UT GT 8.0 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH         A=5   B=4   C=3   D=2   E=1   X=3   RATING    DE/PCT
       U LCC          (VG)  (G)   (A)   (P)   (VP)  (NO)

                      120   583   792   270    75   294   3.21        16
RECORDS  2139
```

```
FIS2 TIME/UNIT SOQ/RATE PCT/AVG LT 75 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH PCT/AVG   A=5   B=4   C=3   D=2   E=1   X=3   RATING  DE/PCT
       U LCC            (VG)  (G)   (A)   (P)   (VP)  (NC)

                        298  1281  1610   281    77   904    3.40       8
RECORDS  4463
```

```
FIS2 TIME/UNIT SOQ/RATE PCT/AVG WL 75,125 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH PCT/AVG   A=5   B=4   C=3   D=2   E=1   X=3   RATING  DE/PCT
       U LCC            (VG)  (G)   (A)   (P)   (VP)  (NC)

                        176   826  1183   328    82   409    3.26      14
RECORDS  3037
```

```
FIS2 TIME/UNIT SOQ/RATE PCT/AVG GT 125 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH PCT/AVG   A=5   B=4   C=3   D=2   E=1   X=3   RATING  DE/PCT
       U LCC            (VG)  (G)   (A)   (P)   (VP)  (NC)

                         70   472   710   285    94   244    3.08      20
RECORDS  1878
```

FIGURE 10

EXAMPLE OF ON-LINE RESEARCH

One of the course development departments produced an PIS II course (70555) which used an experimental adjunct publication.

- The normal adjunct publication for Course #70555 would contain "outline" information only and consist of about 30 pages.

- The experimental adjunct publication contained printed duplicates of all the teaching material presented on the display screens and consisted of about 300 pages.

- Students who complete Course #70555 go to an Education Center for a conventional lecture - laboratory class. Instructors who teach the classes said the incoming students tell them:

  - "...most students study in the experimental publication instead of using the display screens..."

  - "...students like the publication better than the display screens..."

  - "...students using the publication learn faster..."

Course #70555 has 8 units and at 50 student completion records per unit yields a total of 400 student-unit completions for analysis. Two TQRP inquires were used to separate the 400 records into three groups (Figure 11):

- Students who did most of their studying on the terminal and made little use of the experimental publication. These students had off- terminal study time of less than 1 hour (T/OFF LT 1.0). See 1st snd 3rd example reports.

- Students who did most of their studying off the terminal and relied primarily on the experimental publication. These students had off- terminal study time of more than 2 hours (T/OFF GT 2.0). See 2nd and 4th example reports.

- Students who used both the terminal and the experimental publication. These students had an off-terminal study time of between 1 and 2 hours and were not included in the analysis.

```
FIS2 SOQ/RATE COURSE = 70555 T/OFF LT 1.0 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH       A=5  B=4  C=3  D=2  E=1  X=3  RATING   DE/PCT
       U LOC       (VG)  (G)  (A)  (P)  (VP) (NO)

                      7   66  146   23   13   79   3.09      11
RECORDS  334
```

```
FIS2 SOQ/RATE COURSE = 70555 T/OFF GT 2.0 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH       A=5  B=4  C=3  D=2  E=1  X=3  RATING   DE/PCT
       U LOC       (VG)  (G)  (A)  (P)  (VP) (NO)

                           8   12    3         11   3.06       9
RECORDS   34
```

```
FIS2 TIME/UNIT COURSE = 70555 T/OFF LT 1.0 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH PLAN                                 UNIT/AVG    PCT/AVG
       U LOC  TIME

                                                                 84
RECORDS  334
```

```
FIS2 TIME/UNIT COURSE = 70555 T/OFF GT 2.0 T/O.
PG.1 08/18/77        *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH PLAN                                 UNIT/AVG    PCT/AVG
       U LOC  TIME

                                                                206
RECORDS   34
```

FIGURE 11

Now lets compare the report data to the three aforesaid "comments":

- "...most students study in the experimental publication...".

  - Look at the RECORD COUNT for reports 1 and 2.

  - Most students (334) used the terminal. Few used the experimental publication (34) and few used both (400 - 334 - 34 = 32).

  - This comment did not represent the actual situation.

- "...students like the publication better...".

  - Look at the RATING on reports 1 and 2. There is no significant difference in student acceptance (3.09 vs. 3.06) between the two groups.

  - This comment did not represent the actual situation.

- "...students using the experimental publication learn faster...".

  - Look at the PCT/AV on reports 3 and 4 (84% vs. 206%).

  - Students using the terminal complete in 84% of the planned time.

  - Students using the experimental publication complete much more slowly and require 206% of the planned time.

  - This comment did not represent the actual situation.

The three comments resulted from isolated "subjective" opinions. The Course Quality measurement system provided the "objective" analysis that showed the comments to be invalid.

```
FIS2 SOQ/RATE COURSE = 70555 T/OFF LT 1.0 T/O.
PG.1 08/18/77       *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH        A=5   B=4   C=3   D=2   E=1   X=3   RATING    DE/PCT
       U LCC         (VG)  (G)   (A)   (P)   (VP)  (NC)

                      7    66    146   23    13    79    3.09        11
RECORDS  334
```

```
FIS2 SOQ/RATE COURSE = 70555 T/OFF GT 2.0 T/O.
PG.1 08/18/77       *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH        A=5   B=4   C=3   D=2   E=1   X=3   RATING    DE/PCT
       U LOC         (VG)  (G)   (A)   (P)   (VP)  (NO)

                            8    12    3           11    3.06         9
RECORDS   34
```

```
FIS2 TIME/UNIT COURSE = 70535 T/OFF LT 1.0 T/O.
PG.1 08/18/77       *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH PLAN                                      UNIT/AVG    PCT/AVG
       U LOC   TIME

                                                                       84
RECORDS  334
```

```
FIS2 TIME/UNIT COURSE = 70555 T/OFF GT 2.0 T/O.
PG.1 08/18/77       *** FIS2 ANALYSIS REPORT ***
FILE UPDATED ON 08/15/77
COURSE C AUTH PLAN                                      UNIT/AVG    PCT/AVG
       U LCC   TIME

                                                                      206
RECORDS   34
```

FIGURE 11

MEDIA RESEARCH IN PROCESS

The Course Quality measurement system is being used as the measurement vehicle in a two part "media study" now in process attempting to determine:

- Applicability of "media enhancement"
- "Student preference" for types of "complex" media

Part 1: Question: Can a typical "non complex media" unit be enchanced (quality improved) by application of the most appropriate complex media? An existing typical course consisting of 7 units is being used to execute the study. Unit #6 of the course was originally designed to use:

- Display terminal.
- Reference to product service manuals.
- Normal adjunct course publication (course outline).

A review by media experts identified the potential for converting part of the display text and part of the text referenced in manuals to a filmstrip and audio cassette presentation. This part of the study will use the original material for a control group and material "enhanced" with the filmstrip/cassette for the experimental group. Both groups will use the same objectives and test items. The null hypothesis is: "...there will be no difference in Course Quality between the groups...".

Part 2: Question: What type of complex media do students prefer? Unit #3 was originally designed with a filmstrip and audio cassette presentation. A review by media experts indicated that the presentation could be converted to a video presentation and to a minipub* without gain or loss of educational effectiveness. The idea was to determine student preference of various types of complex media.

*Minipub = Pocket novel size publication containing information from the storyboard used in audio/visual presentations.

Three experimental groups will be used. The groups use the same objectives and test items.

- Group 1: Filmstrip and audio cassette
- Group 2: Video cassette
- Group 3: Minipub

The null hypothesis is: "...there will be no difference student acceptance between the groups...". The two other Course Quality components, effectiveness and efficiency will also be measured to ensure that "student preference" does not affect overall quality.

PERFORMANCE ORIENTED TRAINING AND
PEER COACHING IN LARGE CLASSES

BY

JOHN J. KESSLER

AND

STANLEY F. BOLIN

OCTOBER 1977

U. S. ARMY RESEARCH INSTITUTE
ALEXANDRIA, VIRGINIA

Introduction

The Double Horseshoe Method (DHM) was devised to implement the principles of performance-oriented language instruction. The language instruction was required for the improvement of spoken English by Korean soldiers who augment the U.S. Army in Korea. [These soldiers are called KATUSA's, for Korean Augmentation To the U.S. Army. They number approximately 7000 and provide a vital link between U.S. forces and those of the Republic of Korea.] The Double Horseshoe Method is currently employed in all classes at the in-processing center for KATUSA's and in units of the U.S. Second Division.

This paper will describe the nature of the DHM and the way it achieved the goals of performance oriented training. Although the original purpose of DHM was to improve the spoken English ability of soldiers for whom English is a second language, the method could be applicable to a wide range of training topics. The description of the training problem and how the Double Horseshoe Method coped with it will illustrate the potentialities and limits of the method.

The Training Problem

The general problem that was approached in Korea was (and is) a long-standing one: how to teach the KATUSA's spoken English so that they are able to learn military skills in the American Army and be productive soldiers. The constraints on the training situation were severe and allowed only a modest dent in this problem. At the KATUSA Processing Center (KPC) where the developmental work on the DHM was done, Korean soldiers fresh from basic training in the ROK army undergo three weeks of preparation for service in the American army. The median size of a cycle fill has been 240 men. English conversation training is carried out in eighteen two-hour blocks. Normally, six GI instructors are available for each cycle. Their skill in language instruction has been acquired through on-the-job training. They are not bi-lingual.

The students are selected by the ROK army. Serving the three year military obligation as a KATUSA is considered much more desireable than doing it as a ROK soldier. The precise basis of the selection process is unknown other than these men are volunteers and are supposed to have demonstrated the ability to read and write English. According to the Korean school system, middle school graduates receive two years of English instruction and high school graduates as much as an additional four years. A typical class at KPC consists of one-fourth middle school, one-half high school, and one-fourth college graduates. In essence, the training task was to use the eighteen two-hour blocks to orient the students to hearing and making English sounds. Perhaps nowhere was performance oriented training more in order. For nearly all of the students it was the first dialogue with a native speaker.

The Double Horseshoe Method

Doing performance oriented training in classes of 25-45 students was the major training management problem. Previously, the normal complement of six instructors used the lecture method, calling on students from time to time to respond individually or in chorus. Each student had a book which contained a formidable amount of material for which he was held accountable in an end-of-course test.

Observations and independent end-of-course tests by The Army Research Institute-Far East Field Unit (ARI-FE) made it clear that the current training was not effective in improving the ability of the students in spoken English. The principles of FM 21-6, which concerns performance oriented training, were seen as highly applicable; applying them under the circumstances resulted in the development of the DHM.

The DHM took its name from the physical configuration of students in the classroom. They were arranged in two horseshoe-shaped rows with the instructor at the open end. Each student in the inner ring was permanently paired with a student in the outer ring. Several purposes were accomplished by this configuration. One was that the instructor's model behavior was close to the students. His demonstration of the target skills was easily seen and heard. A second was that during the performance by the students of the training material, the instructor attended only to the inner ring. The instructor was thus able to insure that every student performed the desired behavior every time it was required. A third purpose was the demonstration and encouragement of peer coaching. As each student in the inner ring was called on, his partner in the outer ring was ready to help, usually at the direction of the instructor. During the initial class sessions, the instructor demonstrated what and when to coach. After the instructional material was covered with the inner ring of students, the inner and outer rings exchanged seats and the process was repeated. In addition to obtaining the effects of repetition, the procedure contributed to the building of the working relationships between members of the student pairs. This relationship was viewed as essential to promoting practice of the training tasks outside of classroom hours. A further device used to strengthen the peer relationship was the praise or punishment by the instructor of the collective pair.

Scheduling

After the first training session, during which the organization of the class occupied the first hour, a typical two-hour block of training consisted of the following:

1. Test students on material from the previous session, 40 - 50 minutes.

2. Break, 10 minutes

3. Train on new material, 60 - 70 minutes.

Keeping up this schedule was the largest problem encountered in the early stages of DHM development. This was handled in two ways: 1) the amount of material to be covered for each session was cut down to a comfortable time-fit; 2) instructors were told not to spend excessive amounts of class time on individual students.

The importance of adhering to the pattern of the schedule came from several considerations. One was that of insuring that every man would get his chance to perform every behavior under the instructor's supervision. Another was that by separating the training period from the testing period, the illusory gains in ability due purely to immediate memory were eliminated. The time gap also gave peer instruction a chance to occur outside of class. Finally, by requiring a test period after each training period, the instructor was able to gain a quick gauge of the progress (or lack of it) by his students. This allowed the instructor to discover problems almost immediately and to correct them as they became known rather than be surprised at the end of the course.

Results And Discussion

The major question with any training method is how well does it work. As applied to the training of spoken English to KATUSA's, the DHM was a highly successful adaptation of the principles of performance oriented training. Students performed the required training tasks up to the training standards. No experiment was carried out, nor did one seem necessary, to show what was readily observable. There was, however, one small worry. Was there anything lost in switching from conventional instruction to the DHM? The reason for the concern was the large reduction in volume of material that was necessary in changing to performance oriented training. Possibly the better students had been increasing their vocabularies even if they weren't improving their abilities in spoken English.

To determine if this was the case, the pre - and posttraining scores on a written test called the Five Minute English Word Test (5EWT) were examined during the conversion to DHM. The 5EWT required that a student write down as many English words plus their Korean translations as he could in five minutes. In Table 1 are shown the average gain scores for four cycles at the KATUSA Processing Center. Cycle 1 was the first stage in the implementation process and only three of the six instructors were attempting to use DHM. By Cycle 4, all of the instructors were using DHM. That the average gain scores progress steadily as the DHM was implemented is not offered as evidence that the DHM achieves the training goals that were established. These goals were achieved daily in the classroom. Instead, the results in Table 1 show that in spite of reduced and delimited material, no losses were experienced on a measure which would appear to favor the conventional method of instruction previously practiced. The steady increase of gain scores on a written test like 5EWT was a bonus which could have been due to more inspirational instructors and better motivated students as well as the Double Horseshoe Method of training.

Table 1

RESULTS OF FIVE MINUTE ENGLISH WORD
TEST, AVERAGE SCORES BY CYCLE

Cycle	N	Before Training	After Training	Gain	% Gain
7701	241	30.5	36.4	5.9	19.3
7702	244	24.8	31.6	6.8	27.4
7703	200	27.6	36.7	9.1	33.0
7704	99	15.5	25.5	10.0	64.5

Conclusions

The purpose of this paper has been to describe a way of managing performance oriented training in large classes. The Double Horseshoe Method was developed and implemented for the purpose of improving the spoken English ability of Korean soldiers who augment the US Army. When used for this purpose, the DHM was effective on several counts:

1) the G.I. instructors were able to carry it out;

2) student performance on all the material was monitored session by session;

3) peer coaching was demonstrated in the classroom;

4) scores on a measure thought to favor conventional instruction went up during the implementation period;

5) costs were held to the same level.

Using the DHM for subject matter other than spoken English seems quite feasible. The considerations behind its adoption in a particular situation are likely to be similar to those encountered at the KATUSA

Processing Center, i.e., material which lends itself to performance oriented training, an obvious need for performance oriented training, 20 to 40 students per instructor, and perceived benefits from peer coaching. If these conditions are present, the Double Horseshoe Method offers a viable way of managing the delivery of performance oriented training.

References:

FM 21-6, How to Prepare and Conduct Military Training. HQ DA, 3 November 1975

Spoken English Instruction Using the Double Horseshoe Method (DHM), Trainer's Manual. EIGHTH UNITED STATES ARMY KATUSA PROCESSING CENTER, March 1977

POSSIBLE STRATEGIES FOR ESTABLISHING TRAINING PRIORITIES

Arthur C. F. Gilbert, Ph.D.
Raymond O. Waldkoetter, Ed.D.

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

A paper by Waldkoetter et al (1976) reported the utility of the four factor training priority model presented by Mead (1975) in an Army setting. In that paper the efficiency of the four factor model was demonstrated by a multiple correlation of .88 between the four factors and type of training evaluated along a five point scale. The five points on the scale corresponded to five types of training where resident school training carried a value of "5", formal unit training had a value of "4", non-resident courses were assigned a value of "3", on-the-job training had a value of "2", and a value of "1" was assigned in those instances where no training was selected for a task.

The early analysis reported by Waldkoetter et al was based on the assumption that degree of formalization of training could be expressed as a continuous variable extending from a high degree of formalization represented by resident-school training to lack of formalization represented by no training being required. Obviously the two extreme ends of the scale pose no difficulty in terms of definition; however, the ordering of the three intermediate types of training requires certain assumptions about the degree of formalization.

The purpose of this paper is to present possible alternate strategies that might be used in the data analyses that are not based on any assumptions about the extent to which the different types of training are formal in nature.

---

1 The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research or those of the Department of the Army.

PROCEDURE

Waldkoetter et al described the data collection procedure in obtaining the task analysis data on MOS 76V Equipment Storage Specialist tasks. A task list of 183 items was administered to 80 non-commissioned officers and the subjects were required to rate each task on three types of training scales; these scales were Task Learning Difficulty (TLD), Task Delay Tolerance (TDT), and Consequences of Inadequate Performance (CIP). Each scale was a seven-step scale. For the Task Learning Difficulty Scale and for the Consequences of Inadequate Performance Scale, the high end of the scale, "7", corresponded to a high degree of learning difficulty or indicated that the consequences of inadequate performance was great. For the Task Delay Tolerance, a value of "1" indicated that the task should be performance immediately once the need of its performance had been perceived by the incumbent.

Subjects were asked to indicate the type of training that they considered appropriate for each task by indicating whether the task should be taught in resident school, formal unit training, non-resident school training, on-the-job training, or if it did not require any training. For the purpose of analysis in the research by Waldkoetter et al, these five type of training categories were assigned values of "5", "4", "3", "2", and "1" respectively.

As in previous research mean values were derived for each task on the Task Learning Difficulty Scale, the Consequences of Inadequate Performance Scale, and on the Task Delay Tolerance Scale. However, for the analyses reported here the Task Delay Tolerance Scale was reversed in value so that a value of "7" denoted a short delay tolerance. The same data on the percent of members performance each task in the MOS was also used as the fourth factor. Instead of obtaining a mean value on the type of training scale as in the previous research, the frequency of subjects who placed the task in each of the five type of training categories was tabulated for each task.

RESULTS AND DISCUSSION

In Table 1 the correlations among all variables, both predictors and criteria, are shown. Also, in this table the correlations between the five point type of training scale (Waldkoetter et al, 1976) and all of the other variables are shown.

The initial data analysis consisted of deriving the canonical correlations between the set of predictor variables (i.e., Task Learning Difficulty, Task Delay Tolerance, Consequences of Inadequate Performance, and Percent of Members Performing) and the set of type of training criterion variables (i.e., Resident School Training, Formal Unit Training, Non-Resident School Training, On-the-Job Training, and No Training). Three sets of canonical variates were derived yielding canonical correlations of .91, .49, and .86 all of which were significant beyond the .01 level.

Table 1

CORRELATIONS AMONG THE TYPE OF TRAINING VARIABLES AND THE FOUR PREDICTOR VARIABLES

Variable	TLD	CIP	TDT	PCT	TOT	NT	OJT	NRS	FUR	RST
TLD	1.00	.56	.11	.-.39	.87	-.74	-.76	.57	.20	.85
CIP		1.00	.59	-.16	.52	-.57	-.39	.30	.30	.45
TDT			1.00	.17	.12	-.24	-.05	.17	.19	.04
PCT				1.00	-.38	.30	.40	-.20	-.19	-.37
TOT					1.00	-.78	-.87	.44	.37	.96
NT						1.00	.42	-.39	-.39	-.65
OJT							1.00	-.51	-.27	-.89
NRS								1.00	-.17	.43
FUT									1.00	.15
RST										1.00

[1]TLD = Task Learning Difficulty
CIP = Consequences of Inadequate Performance
TDT = Task Delay Tolerance
PCT = Percent of Incumbents Performing
TOT = Type of Training (5-point scale)
NT = No Training Required
OJT = On-the-Job Training
NRS = Non-Resident School Training
FUT = Formal Unit Training
RST = Resident School Training

The first set of canonical variates indicated that tasks judged high in Task Learning Difficulty tended to be perceived as requiring resident school training. The second set of variates revealed that tasks that were judged as being difficult to learn and which a large percentage of members perform tended to be associated with resident school training or on-the-job training. There was some tendency to view tasks in which the Consequences of Inadequate Performance were considered high as not being appropriate for non-resident school training. The third set of canonical variates tended to indicate that for those tasks in which a low percent of incumbents perform and when the task delay tolerance is not important neither on-the-job training or resident school training is perceived as being appropriate.

A factor analysis was performed for the four predictor variables Two factors emerged in this analysis. The first factor was identified by Task Delay Tolerance and Consequences of Inadequate Performance while the second factor was bipolar. At one end it was identified by task learning difficulty and at the other by percent of members performing.

The second factor analysis involved the set of criterion variables (or type of training variables). Two factors were also extracted from this set. Both of these factors were bipolar. The first factor was identified at one end by resident school training and at the other by on-the-job training. Both non-resident school training and the no training variables loaded substantially (i.e., above an absolute value of .40) on this first factor. The second bipolar factor was identified at the positive end by formal unit training and at the other end by no training being required and by non-resident school training.

Five regression analyses were performed using each of the five type of training variables as the criterion. In each analysis all of the four predictor variables were used. The beta weights derived in each of these analyses and the corresponding multiple correlation and squared multiple correlations are shown in Table 2. For the sake of comparison the same data are shown for the five point type of training scale reported by Waldkoetter <u>et</u> <u>al</u>. Four of the multiple correlations were significant at the .01 level, while the fifth predicting formal unit training was significant at the .05 level. The multiple correlations ranged from .8538 for resident-school training to .3367 for formal unit training.

The final analysis of the data involved applying the regression weights in the five analyses of regression and the weights derived from the Waldkoetter <u>et</u> <u>al</u> analysis to obtain a predicted type of training for each task. The correlations among the six predicted types of training were then computed across the 183 tasks. The matrix of correlations resulting from these computations are shown in Table 3.

Table 2

BETA WEIGHTS OF THE FOUR VARIABLES IN THE FIVE REGRESSION EQUATIONS FOR PREDICTING TRAINING PRIORITIES

	Type of Training 5-point Scale	No Training Required	OJT	Non-Resident School Training	Formal Unit Training	Resident School Training
Task Learning Difficulty	.8309	-.6113	-.7327	.6392	-.0006	.8402
Consequences of Inadequate Performance	.0503	-.1780	.0426	-.1745	.2182	.0055
Task Delay Tolerance	.0024	-.0742	-.0160	.2022	.0908	-.0517
Percent of Members Performing	-.0489	.0481	.1240	-.0167	-.1665	-.0357
R	.8802	.7644	.7678	.5909	.3367	.8538
R^2	.7747	.5843	.5896	.3492	.1134	.7290

Table 3

CORRELATIONS AMONG THE PREDICTED TYPES OF TRAINING

Variable[1]	TOT	NT	OUT	NRS	FUT	RST
TOT	1.00	-.98	-.99	.96	.63	.99
NT		1.00	.94	-.94	-.76	-.95
OJT			1.00	-.95	-.61	-.99
NRS				1.00	.58	.95
FUT					1.00	.57
RST						1.00

[1]TOT = Type of Training (5-point scale)

NT = No Training Required

OUT = On-the-Job Training

NRS = Non-Resident School Training

FUT = Formal Unit Training

RST = Resident School Training

The results of the canonical correlation analysis indicated the trends that exist in the data in highlighting the relationship that exists between difficulty in the learning of a task and its appropriateness for resident school training. Another finding of interest from this analysis centers on the fact that if a task is considered as being difficult to learn and if it is being performed by a large percent of incumbents then it is viewed as being appropriate for either resident school training or on-the-job training.

Factor analysis of the type of training variables indicated two, instead of a single factor emerging, thus showing the possibility of ordering these variables along a single dimension may not be entirely valid. However, it should be pointed out that four of the five training categories loaded highest on the first factor, the exception being formal unit training which defined the positive end of the second factor.

The results of the regression analyses indicated that the four factors were most efficient in predicting resident school training. The resulting multiple correlation was .85 compared with the multiple correlation of .88 obtained for the five-point type of training scale. The lowest multiple correlation was obtained in predicting formal unit training. This could very well be due to the lack of a common definition of this variable among the raters. Again, perhaps for the same reason, the next lowest multiple correlation was obtained when non-resident school training was used as the criterion.

Examination of the matrix showing the correlations among the different predicted types of training (shown in Table 3) reveals that, generally, the correlation between the predicted formal unit training and other predicted types of training is lowest. The correlations in this table support the concept that whatever strategy is employed in selecting a criterion against which to validate the four factor model, the utility of the model in defining training priorities is upheld with this one exception.

REFERENCES

Mead, D. F. Determining training priorities for job tasks. Paper presented at the 17th Annual Conference of the Military Testing Association, Indianapolis, Ind., September 16-19, 1975. In Proceedings of the 17th Annual Conference of the Military Testing Association. Indianapolis, Ind.: U.S. Army Enlisted Evaluation Center, 1975.

Waldkoetter, R. O., Gilbert, A. C. F., Raney, J. L. and Hawkins, H. H. Establishing training priorties in an Army Supply Military Occupational Specialty. Paper presented at the 18th Annual Conference of the Military Testing Association. In Proceedings of the 18th Annual Conference of the Military Testing Association. Pensacola, Fla.: Naval Education and Training Program Development Center, 1976.

UTILIZATION OF DIFFERENTIAL PROFICIENCY LEVELS FOR CRITERION-REFERENCED TRAINING SYSTEM ASSESSMENT

John B. Meredith, Jr.
Data-Design Laboratories
Norfolk, Va. 23505

INTRODUCTION

In designing criterion-referenced, multiple-choice tests, one of the most perplexing problems is the determination of the passing score. Either passing scores are arbitrarily set at some percent correct (Sanders, 1976; Shaycoft, 1976) or they are determined by complex mathematical methods that incorporate α and β classification errors (Emrick, 1971; Kriewall, 1972; Hively et al, 1973; Millman, 1972 and 1973; Roundabush, 1974; and Wilcox, 1977). An α classification error occurs when a nonmaster is falsely deemed to be a master; conversely a β classification error occurs when a master is falsely deemed to be a nonmaster (Meskauskas, 1976). In most of the previous studies, the methods for determining passing scores were either too simplistic, thereby, resulting in large classification errors (Reichman and Oosterhoff, 1976; or requiring complex parameter estimation procedures (Wilcox and Harris, 1977).

In this study an approach for determining passing scores developed by Nedelsky (1954) is utilized. This approach involves the use of Subject Matter Experts (SMEs) to define the minimal performance level for a test by rating the difficulty of each alternative to each test item for the minimally acceptable (just passing) examinee (Meskauskas, 1976). Thus, this procedure establishes item content rather than examinee performance as the basis for determining item difficulty (Smilansky and Guerin, 1976). This approach for setting passing scores provides one of the best estimates of the probability of classifying examinees into correct mastery or nonmastery states (Reichman and Oosterhoff, 1976).

When evaluating a heterogeneous group of personnel with varying experience, the comparison of experienced personnel to inexperienced personnel with a single passing score (proficiency level) is not appropriate. This type of comparison would pose a serious threat to the external validity by questioning the generalizability of the results to the entire population (Bracht and Glass, 1968).

To overcome this external threat to the validity of the findings, multiple passing scores can be determined according to various levels of personnel experience. In this study, experience was defined at three levels based upon personnel watch station qualifications. These three levels were:

1. Apprentice Technician (replacement school graduate)

2. Journeyman Technician (qualified watchstander)

3. Master Technician (qualified watch supervisor)

From these multiple proficiency levels, Minimally Acceptable Performance Levels (MAPLs), were determined by proficiency level for each area of a test. The purpose of this study was to determine the extent to which replacement and advanced training curricula produced effectively trained technicians.

METHOD

Nedelsky (1954) developed the technique as an "absolute standard" for evaluation of physics students on a departmental, multiple-choice comprehensive examination at the University of Chicago. The technique was validated by Taylor and Reid (1972), Bobula (1974), Smilansky and Guerin (1976), and Meredith (1977). The use of this technique is dependent upon the assumption that SMEs can define alternative similarity as follows:

1. An alternative which a minimally acceptable examinee <u>should</u> recognize as incorrect is given a value of <u>zero</u> (0).

2. An alternative which a minimally acceptable examinee <u>should not</u> recognize as incorrect is given a value <u>of two</u> (2).

3. An alternative which is correct is given a value of two (2).

4. All other alternatives are given values of one (1).

An example of the application of this technique, adapted from Bobula (1974), is exemplified by the faculty member who is teaching statistics and defines ability to recognize measures

of central tendency as a component of basic statistical competency. He might write an item and assign the following values:

	Value	
		"The most appropriate measure of central tendency for the reading performance scores of a large class is the ..."
	(2)	1. Mode
	(0)	2. Variance
Correct	(2)	3. Mean
	(2)	4. Median
	(0)	5. Standard Deviation

When the SME rated this item for the "minimally acceptable" examinee, he decided that three alternatives (Mean, Median, and Mode) were equally viable, and all were given values of two (2). The SME rated two alternatives (Variance and Standard Deviation) as not meeting the minimal component of basic statistical competency, and both were given values of zero (0).

The Alternative Similarity Index (ASI) for the i^{th} multiple-choice test item at a given proficiency level is defined as follows:

$$ASI_i = \frac{2n}{\sum_{j=1}^{n} \sum_{K=1}^{m} A} \qquad (1)$$

Where: n = number of SMEs,

m = number of alternatives, and

A = value assigned to m^{th} alternative by the n^{th} SME.

In the preceding statistical example, the ASI was calculated by equation (1) as follows:

$$\text{ASI} = \frac{2(1)}{\sum_{j=1}^{1} \sum_{k=1}^{5} A} = \frac{2}{2+0+2+2+0} = \frac{2}{6} = .33$$

The MAPL for each area of specialization is determined by summing the ASIs at a given proficiency level over the number of items within the area. For each area of specialization on the test, there will be three MAPLs, one at each proficiency level. Any technician at a given proficiency level, whose area raw score exceeds the respective MAPL, is deemed to be a master; conversely, any technician at a given proficiency level, whose area raw score does not exceed the respective MAPL, is deemed to be a nonmaster. A technician may only be considered a master/nonmaster at one proficiency level. That is, technician watch station classifications are independent and mutually exclusive.

At each proficiency level, the technician's raw scores were transformed to standardized area scores. This transformation was appropriate in order to meet the assumptions of an analysis of variance design (Cochran and Cox, 1957). The transformation was based upon a standardized area criterion score ($Z_{i,j}$) for each technician and is defined as follows:

$$Z_{i,j} = \frac{X_{i,j} - \text{MAPL}_{i,j}}{S_i} \qquad (2)$$

Where: $X_{i,j}$ = raw examinee score for i_{th} area, j^{th} proficiency level.

$\text{MAPL}_{i,j}$ = MAPL score for i_{th} area, j^{th} proficiency level.

S_i = standard deviation of raw scores for i^{th} area.

From standardized area criterion scores, a factorial analysis of variance was utilized to test differences among areas and proficiency levels.

This analytical design was a pure Model I, where both effects (areas and proficiency levels) were fixed. This is represented by (Sokal and Rohlf, 1969):

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Where: Y_{ijk} = k^{th} technician representing the i^{th} area and the j^{th} proficiency level.

μ = parametric mean of technician population.

α_i = fixed treatment effect for i^{th} area.

β_j = fixed treatment effect for j^{th} proficiency level.

$(\alpha\beta)_{ij}$ = interaction effect between i^{th} area and j^{th} proficiency level.

ε_{ijk} = error term of k^{th} technician at the j^{th} proficiency level for the i^{th} area.

The null hypothesis of interest in this study was that there would be no significant difference among areas.

$$H_0: \mu_a = \mu_b = \ldots = \mu_l$$

When a significant difference is found among areas, a post hoc, multiple comparison test is utilized to determine which area(s) differ from the other areas. Any area that is significantly below the other areas may indicate undertraining. Conversely, any area that is significantly above the other areas may indicate overtraining.

FINDINGS

In this study, a total of 17 SMEs evaluated a 260-item test that was composed of 12 areas of specialization (A through L). Each SME evaluation was used only at one proficiency level. In that, 12 SMEs were used to determine the Apprentice Technicians' MAPL for each area, 2 SMEs were used to determine the Journeyman Technicians' MAPL for each area, and 3 SMEs were used to determine the Master Technicians' MAPL for each area.

This test was administered to Navy submarine technicians (N=317). The technicians were subsequently divided into

three, mutually exclusive, watch station categories. From each of the three watch station categories, 28 technicians were randomly selected and utilized in the analysis. The number of technicians randomly selected was based upon the power of the analysis ($1-\beta$) and was determined to be .95 (Cochran and Cox, 1957).

The area MAPLs by proficiency level are presented in Table 1.

Table 1.

Area MAPLs by Proficiency Level

Area	# Items In Area	Apprentice Technician	Journeyman Technician	Master Technician
A	11	4.994	5.577	6.908
B	18	8.478	11.124	11.430
C	18	9.612	10.350	11.250
D	41	21.279	23.944	27.429
E	26	13.936	13.936	16.458
F	20	10.700	10.700	13.160
G	20	11.160	11.160	13.000
H	20	11.000	11.000	11.000
I	21	9.870	11.340	13.230
J	21	11.067	11.067	12.810
K	18	9.774	9.774	10.296
L	26	16.302	16.302	17.082

NOTE: For purposes of this study, an additional constraint was imposed. This constraint was that neither Journeymen nor Master Technicians' area MAPL may be lower than the Apprentice Technician MAPL. Also, Master Technicians' area MAPL may not be lower than the Journeyman Technician area MAPL.

The null hypothesis of no difference among areas was rejected. The analysis of variance summary is presented in Table 2.

Table 2.

Analysis of Variance Summary

Source Variation	Sums of Squares	Degrees of Freedom	Expected Mean Square	Mean Square	F-Value	Level of Significance
Among Areas	160.420	11	$\sigma^2 + Nb\,\frac{\Sigma\alpha^2}{a-1}$	14.584	16.331	.000000006
Among Watch-Stations	1.828	2	$\sigma^2 + Na\,\frac{\Sigma\beta^2}{b-1}$	.914	1.024	.3595
Areas X Watch-Stations	63.599	22	$\sigma^2 + N\Sigma\,\frac{(\alpha\beta)^2}{ab-a-b+1}$	2.891	3.237	.0000008
Error	868.350	972	σ^2	.893		
Total	1,094.197	1,007				

A post hoc, multiple comparison test (Newman-Keuls) was performed to determine which area(s) differed significantly from the other areas. This was appropriate since the area effects from the analysis of variance were highly significant. Since a highly significant interaction between areas and watchstations was also found, the interpretations of the area effects become more complex. With a highly significant interaction effects, global training assessments based upon area effects alone would probably be misleading. Therefore, graphical analyses of mean area scores are presented in Figures 1, 2, 3, and 4 to facilitate interpretations of the results (Winer, 1971).

From the graphical analyses, the following findings are reported by area:

1. In area A, all technician groups were above expectations. (This area dealt with casualty procedures which are very thoroughly taught.)

2. In area B, apprentice technicians performed significantly higher than expected when compared to the other technician groups. This may indicate the possibility of overtraining in this area for initial replacement training.

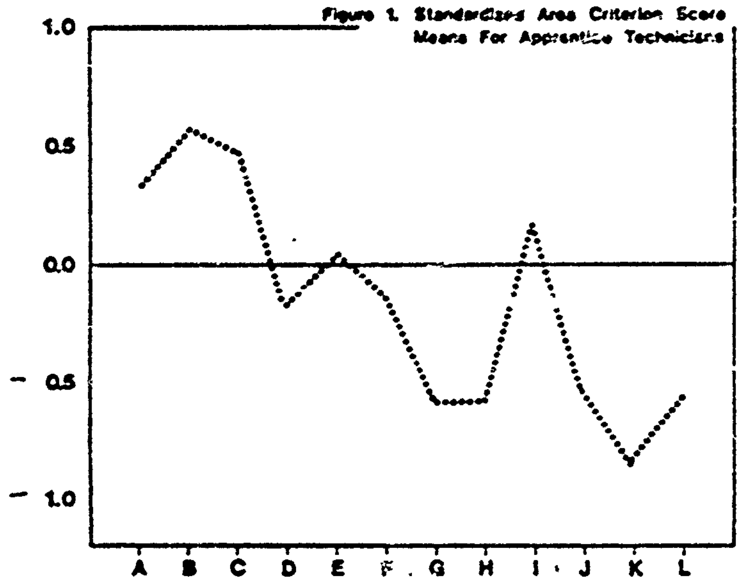


Figure 1. Standardized Area Criterion Score Means For Apprentice Technicians

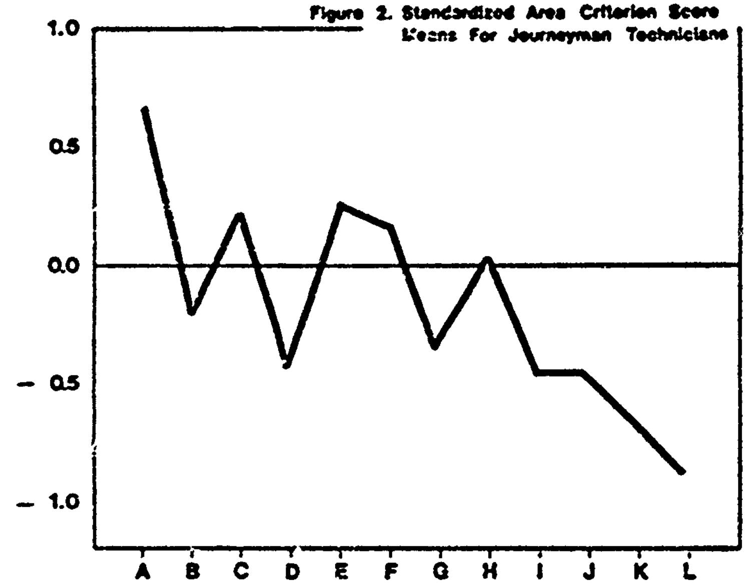


Figure 2. Standardized Area Criterion Score Means For Journeyman Technicians

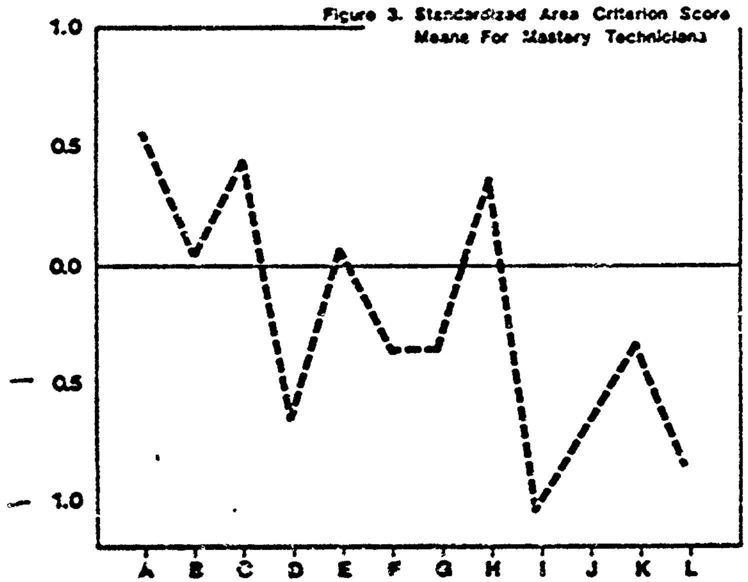


Figure 3. Standardized Area Criterion Score Means For Mastery Technicians

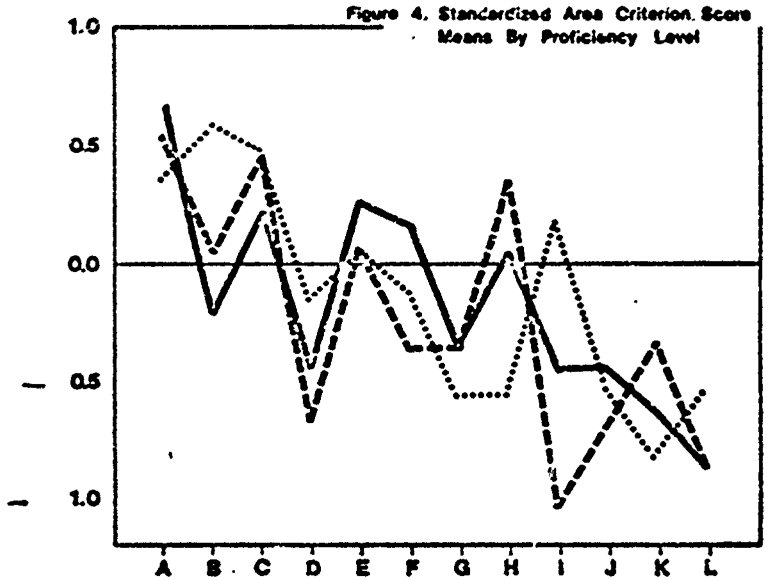


Figure 4. Standardized Area Criterion Score Means By Proficiency Level

3. In area C, all technician groups performed somewhat above expectations. This gives an indication that technician performance is satisfactory (above expectations).

4. In area D, all technician groups performed somewhat below expectations. However, master technicians performed significantly below expectations. This may indicate that refresher courses (advanced training) may be needed for master technicians.

5. In area E, all technicians groups consistently performed as expected.

6. In area F, all technician groups performed adequately. However, a large variance among technician groups requires further investigation.

7. In area G, all technician groups performed below the expected performance levels. The apprentice technicians performed at such a level which may indicate a lack of training.

8. In area H, journeyman and master technician groups performed as expected. The apprentice technician group performed at such a level which may indicate undertraining.

9. In area I, apprentice technicians performed as expected with the journeyman technician group performing somewhat below expectations. The master technician group, however, performed significantly below expectations which may indicate a serious undertraining problem.

10. In area J, all technician groups performed below expectations which may indicate a general undertraining trend in this area of specialization.

11. In area K, master technicians performed somewhat below expectations. The apprentice and journeyman technician groups, however, performed significantly below expectations which may indicate a general undertraining in this area of specialization.

12. In area L, all technician groups performed significantly below expectations. This may indicate serious undertraining in this area of specialization.

CONCLUSIONS

This paper has extended a methodology for making preliminary training assessments based upon criterion-referenced, multiple-choice tests from the use of a single proficiency level (Meredith, 1977) to multiple proficiency levels. The extension to multiple passing scores has resulted in a more penetrating assessment of training because technicians of varying experience were evaluated at their experience level. To compare experienced technicians to a criterion based upon unexperienced technicians or vice-versa may be inappropriate. With the use of multiple proficiency levels, evaluators are able to fine tune a training system to the specific needs of various levels of technicians.

The limitation of the multiple MAPL procedure is that additional number of SMEs are required for evaluating the items in multiple-choice tests.

One solution to this problem is to have each SME evaluate items for more than one technician proficiency level. In this study, this was determined to be inappropriate due to a carry over effect from one proficiency level to another by SMEs, thus imposing an additional assumption of local independence among SMEs.

The education implications of this study may be in the determination of minimal performance criteria (passing scores) for a test or test part. This may be extremely useful for those of us involved with competency-based education. In particular, educators could evaluate the effectiveness of their system at various points in the curriculum while at the same time evaluating the product (students). In this manner, educators could receive quantitative data from which adjustments could be made in the relative emphasis of their programs. This could result in a better attuned system that will meet the needs of the students and will result in a more efficient allocation of limited resources.

REFERENCES

1. Bobula, J. Minimum Pass Level Study. Report to the Faculty, Center for Educational Development, College of Medicine, University of Illinois, 25-31, 1974.

2. Bracht, G. and Glass, G. The External Validity of Experiments. American Educational Research Journal, 5:437-74, 1968.

3. Cochran, W. G. and Cox, G. M. Experimental Design, New York, John Wiley and Sons, 1957.

4. Emrick, J. A. An Evaluation Model for Mastery Testing. Journal of Educational Measurement, 8:321-6, 1971.

5. Hively, W., Maxwell, G., Rabehl, G., Sension, D. and Lundin, S. Domain - Referenced Curriculum Evaluation: A Technical Handbook And A Case Study from the MINNEMAST Project. UCLA Graduate School of Education, Center for the Study of Evaluation, 1973.

6. Kriewall, T. E. Aspects and Applications of Criterion-Referenced Tests. Institute for Educational Research, April, 1972. (ERIC Document Reproduction Service No. ED 063 333, 27 pp.)

7. Millman, J. Determining Test Length, Passing Scores and Test Lengths for Objective-Based Tests. Los Angeles: Institutional Objectives Exchange, 1972.

8. Millman, J. Passing Scores and Test Lengths for Domain-Referenced Measures. Review of Educational Research, 43:205-16, 1973. (ERIC Document Reproduction Service No. ED 065 555, 17 pp.)

9. Meredith, Jr. J. B. The Use of A Criterion-Referenced Test Methodology to Assess Training System Effectiveness Paper presented at The American Educational Research Association, New York, New York, April 1977.

10. Meskauskas, J. A. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard Setting, Review of Educational Research, 46:133-158, 1976.

11. Nedelsky, L. Absolute Grading Standards for Objective Tests. Educational and Psychological Measurement, 14:3-19, 1954.

12. Reichman, S. L. & Oosterhoff, A. C. Strategy Guidelines for the Construction of Mastery Tests, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April, 1976.

13. Roundabush, G. E. Models for a Beginning Theory of Criterion-Referenced Tests. Paper presented at the Annual Meeting of the National Council for Measurement in Education, Chicago, April, 1974.

14. Sanders, J. R. Alternatives for Achievement Testing, Educational Technology, 16:March, 1976.

15. Shaycoft, M. F. A Paradox in Setting Cutting Scores on Criterion-References Tests, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April, 1976.

16. Smilansky, J. and Guerin, R. O. Minimal Acceptable Performance Levels for Criterion-Referenced Multiple-Choice Examinations and Their Validation, Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, California, April, 1976.

17. Sokal, R. R. and Rohlf, F. J. Biometry, San Francisco, W. H. Freeman and Co., 1969.

18. Taylor, D. D. & Reid, J. C. Criterion-Referenced Score of a Mastery Test, Journal of Medical Education, 47:970-971, 1972.

19. Wilcox, R. R. A Note on the Length and Passing Score of a Mastery Test, Journal of Educational Statistics, 4:359-64, 1977.

20. Wilcox, R. R. and Harris, C. W. On Emrick's "An Evaluation Model for Mastery Testing," Journal of Educational Measurement 14:215-218, 1977.

21. Winer, B. J. Statistical Principles in Experimental Design, New York, McGraw-Hill, 1971.

Relating Task Surveys to the Content of Existing Training Programs

Harry L. Ammerman, Ph.D.
The Center for Vocational Education
Columbus, Ohio

During the many years that a number of us have conducted research on military technical training and testing, we often realized our results also had potential for civilian applications. Now I have the opportunity to reverse that circumstance, by describing a methodology developed in the public education sector that may have a usefulness in military contexts. I hope you find it of interest and value.

It is taken for granted that most of you are aware of the recent emphasis upon the validation of employee selection tests, as put forth in Federal Executive Agency Guidelines. Related to this emphasis, the 1976 federal court decision in the case of Washington vs. Davis begins to highlight a concern for the content validity of training programs, potentially subjecting curriculum content (as well as the content of achievement tests based on that training) to standards of validation comparable to those standards being imposed upon employee selection tests.

The basis for such validation of training content already exists for many of you in your task inventories and surveys of occupational performance. By these means, and by related methods used in the engineering of instructional systems, there generally is produced an identification of what work tasks are relevant to a defined occupation. There also is some selection of which tasks are appropriate for formal school training, as well as some specification of what task content and performance standards are to be of concern in school training for each occupation.

A problem can arise at this point when the results of these front-end analyses are to be compared to the content of an existing training program, as might be done to see if there are any significant discrepancies between the two. This comparison is not too difficult when the content of the training program is given in the same terms as the occupational task survey results; that is, in terms of specific tasks and of the knowledges, skills, and proficiencies associated with each task.

However, an existing curriculum may happen to be stated in a somewhat different form, making comparisons very difficult. A particular program might state its training content in such other forms as these:

1. Listing or outlining the topics, concepts, generalizations, system components, or other such elements of knowledge content to be dealt with.

2. Descriptions of what instructors are to do, or the activities to be carried on by the teachers.

3. Stated as highly generalized patterns of behavior, noting the general kinds of changes in students with which the program is intended to deal.

4. Student performance objectives stated in terms of school-related behavior.

5. Listing of the titles of available instructional courses, as in a college catalog.

6. Identification of particular textbooks, student workbooks, teaching aids, laboratory exercises, and other instructional resources to be used.

7. Test items which convey the intended areas of ability development or learning attainment.

When an existing curriculum is described only in one or more of these other ways, the relation of that curriculum to job performance content derived by task survey procedures is not readily apparent. It would be helpful if present curriculum content were convertable to a form more similar to occupational survey results.

Defining Curriculum Content

Let me pause at this point to define what is meant here by the term "curriculum content," and to suggest the key variables to be used in identifying curriculum content.

The concept of "curriculum" is considered to mean the "intended learning outcomes" that have been selected and ordered. This view of curriculum as being a product that states "what is to be learned" by a student is based upon Mauritz Johnson's 1967 and 1969 definitive considerations in curriculum theory, wherein he distinguishes between the concepts of "curriculum" and "instruction."

Instruction becomes the process by which intended learning outcomes are achieved. "Curriculum," on the other hand, is the planned and structured series of those intended learning outcomes. Thus, the basic distinction is made here between (a) what is to be learned (that is, the curriculum) and (b) how such learning is to be attained (the instruction).

Curriculum content" is identified on the basis both of its intended inclusion and its emphasis in a training program. The use of content inclusion and emphasis as the key variables of curriculum content is based on the conclusions of Decker Walker and Jon Schaffarzick from their extensive study in 1974 of what are the important influences on student learning achievement.

These features of content inclusion and content emphasis are what we have attempted to operationalize. This has been done in an instrument to be applied to an existing training program for the purpose of identifying what is the job performance content that constitutes the planned and intended outcomes of that training. The instrument, patterned as a task questionnaire, identifies the intended curriculum content in a manner that improves our ability to make direct comparison with task survey results or other such front-end analyses resulting from instructional system design efforts.

"Content inclusion," as that concept is operationalized here, is concerned with whether each particular task of an occupation is or is not intended to receive some consideration in the training program. "Content emphasis" is concerned with the level of development of performance ability that is intended in the school training. This indicates DEGREE of task emphasis.

Additionally, "content emphasis" can pertain to AREA of task emphasis, where particular non-performance features of a task are especially important and attended to in the training process.

Curriculum Content Questionnaire

These three identifiers of curriculum content form the basis for task questions used in a Curriculum Content Questionnaire that is administered to school personnel who are most knowledgeable about what learning is intended in a particular training program. This questionnaire is the means by which curriculum representatives who are knowledgeable of planned program content may indicate the nature and emphasis of job content existing in a curriculum.

The questionnaire is similar in format to a Task Inventory Questionnaire.

Listing of Tasks	Question 1 Level of Ability Intended To Be Developed (Circle One Level)	Question 2 Task Areas To Be Emphasized (Enter Number of Area)
1.	0 1 2 3 4 5 6 7	______
2.	0 1 2 3 4 5 6 7	______
3.	0 1 2 3 4 5 6 7	______
⋮	⋮	⋮

Figure 1. Format of Curriculum Content Questionnaire

As shown in Figure 1, it consists of a column listing the tasks of an occupation, followed by two special task questions. The first question seeks to identify both task inclusion in training and its degree of emphasis with respect to the level of task ability. The second question probes for task areas that are especially important and intended to be emphasized in the training relevant to each task.

The Appendix contains complete directions and explanation of the response categories for each task question. Here let me show abbreviated versions of each response scale.

<u>Level of Task Development</u>

The first question asks each respondent to rate the extent that the curriculum, during the training program, deliberately plans to develop task proficiency. Eight levels of task development are possible, with the levels being defined in a manner similar to that used by John Hemphill (1960) on his scaling of the job significance of tasks:

0 NO DEVELOPMENT of the task is intended

1 Develop only a GENERAL AWARENESS of the task

2

3

4 Develop a BASIC ABILITY to perform the task

5

6

7 Develop a VERY HIGH PROFICIENCY in the skillful performance of the task

Responses other than "0" would be used to indicate some degree of inclusion of a task in the planned learning. This inclusion could then range from a minimal general awareness of the task to development of very high proficiency in performing the task. The midpoint of level "4" depicts a basic ability to do the task, but implies no special intent that any advanced speed, accuracy, or excellence of task performance be developed. This is typically the most frequently used category, with level "7" being next most frequent. Levels higher than "4" represent more advanced levels of skill development with increasingly higher standards of speed, accuracy, or excellence of task performance.

Table 1 portrays the percent of times raters used each response category when the Curriculum Content Questionnaire was applied to training programs in three different occupational areas.

Table 1

Percent of Category Usage
On Level of Development Scale

Scale Category	Training Program: Mechanic	Programmer	Secretary
0 - NO DEVELOPMENT	12%	19%	11%
1 - GENERAL AWARENESS	[illegible]	11	6
2	6	4	5
3	4	6	6
4 - BASIC ABILITY TO DO	18	21	27
5	11	7	17
6	14	10	13
7 - VERY HIGH PROFICIENCY	24	16	15
Missing Data	1	4	0

Nine raters were used for each type of training program, with 180 tasks rated in each occupation. This type of scale helps distinguish between different levels of task proficiency by stretching ratings of developed task performance over four categories, levels "4" through "7". Ratings for each task are summarized across raters by taking the mean value of the 0 - 7 scale. Mean values above 3.0 generally indicate that some amount of task training is in fact planned for the curriculum. However, no one precise mean rating was found to accurately designate the point that differentiated between "no development" and "some development" intended in training. This distinction appears to be identified better by other means, using survey data from workers and supervisors on questions of task occurrence and significance. However, the higher the value for level of development, the more likely the intent to include the task in training.

Interrater reliability for the nine judges over 180 tasks in each occupation was .88 for Mechanics, .78 for Programmers, and .83 for Secretaries. These reliabilities are adjusted for mean differences among raters, and were calculated using Ben Winer's analysis of variance procedure (1971).

It is of interest to note the obtained relationship between this level of development scale and other questions often used on Task Inventory Questionnaires. We had some task survey data available from 120 workers in each occupation. Their mean responses were correlated against mean levels of task development, as shown in Table 2.

As can be seen in this table, task survey data for Mechanics tended to correlate quite high with intended levels of development. However, for the less prescribed, less routine types of occupations, these correlations dropped off considerably, though all retained statistical significance.

Table 2

Correlations Between Task Survey Data And Level of Development Scale

	Training Program		
Task Survey Measure	Mechanic	Programmer	Secretary
. Percent of Workers Performing Each Task	.89	.76	.61
. Frequency Of Task Performance: Based On All Workers Surveyed	.83	.76	.63
Based Only On Workers Who Perform The Task	.32	.56	.58
. Relative Proportion Of Time Spent	.77	.30	.41
. Extent Task Is Part Of The Job (Hemphill scale of significance)	.91	.72	.59

Areas of Emphasis

Now, turning to the second question on the Curriculum Content Questionnaire, we have postulated 11 task areas that might be especially important for emphasis in task training. These areas represent matters other than performance speed and accuracy, since such features are already identified by the level of development question.

Four of the areas pertain to various aspects of the job content; two pertain to personal matters, while the remaining five pertain to technical knowledge and skill areas. Each of these is accompanied by a brief definition in the questionnaire (see Appendix). Raters are asked to indicate which, if any, areas other than performance ability are especially emphasized in training for each task.

Usage of the several categories of emphasis varied among the three occupations in which we administered the Curriculum Content Questionnaire, though Technical Knowledge (Category 9) predominated.

The distribution in Table 3 again reflects nine raters and 180 tasks per occupational area.

Table 3

Percent Of Category Usage
For Areas Of Training Emphasis

	Training Program		
Area Of Emphasis	Mechanic	Programmer	Secretary
1 - Order, Timing	9%	10%	12%
2 - Value Purpose	12	18	4
3 - Safety	8	1	2
4 - Varied Conditions	5	1	9
5 - Relating To Others	3	4	11
6 - Attitude, Responsibility	9	10	17
7 - Basic Education	4	8	6
8 - Detect Discrepancies	15	7	8
9 - Technical Knowledge	18	25	24
10 - Job Aids	11	8	5
11 - Alternate Methods	5	8	3

The figures showing 10% or more of the responses are highlighted by circles on this table. Sensitivity to the value and importance of tasks was emphasized more often for Mechanics and Programmers than for Secretaries, whereas, secretarial training more often emphasized the development of pride in work done and their feelings toward doing quality work, though I hasten to add that these matters were not totally neglected for the other two occupations. Raters of Mechanic training programs tended to mark twice as many areas for emphasis than did raters for the other two occupations.

These patterns of actual distributions of rater use of the emphasis categories are modified somewhat when the responses are summarized across nine raters.

As it turned out, not all 180 tasks per occupation were intended for training, leaving some for on-the-job experience and learning. Areas for each task were summarized by noting simply where four or more of the nine raters agreed on the same area of emphasis. The results are shown in Table 4.

In the distribution that results from looking only where such a level of agreement existed, we find that Mechanics increased their proportions for task Value and Safety, as well as for Detecting Discrepancies and for Technical Knowledge. All other areas decreased their proportions in comparison to raw frequencies of category usage. For Programmers, areas of Worker Responsibility and Technical Knowledge increased their proportions, with other areas decreasing. For Secretaries the increases in proportion were evident in the areas of Relating to Others and Technical Knowledge.

As an aside, it was interesting to note that when a comparable question was asked of employers regarding their expectations for trained graduates, (Ammerman & Essex, 1977), the Basic Education area was more evident for Mechanics and Secretaries. The Mechanic areas of Safety, Detecting Discrepancies, and use of Job Aids were much less evident in employer expectations, as compared to training intentions. However, interpreting such differences between training intentions and employer expectations goes beyond the scope of this paper.

Table 4

Number Of Tasks Per
Area Of Training Emphasis

Area Of Emphasis	Training Program Mechanic	Programmer	Secretary
1 - Order, Timing	1	.	4
2 - Value, Purpose	17	3	.
3 - Safety	17	.	.
4 - Varied Conditions	.	.	.
5 - Relating To Others	1	.	10
6 - Attitude, Responsibility	6	3	9
7 - Basic Education	.	.	1
8 - Detect Discrepancies	36	1	3
9 - Technical Knowledge	53	16	31
10 - Job Aids	10	.	.
11 - Alternate Methods	1	1	.
Emphasis Totals	142	24	58

Summarizing The Results To Show Training Intentions

The results from the two questions of the Curriculum Content Questionnaire, as applied to three different types of occupational training programs that precede employment in those jobs, are illustrated in Tables 5 and 6.

Table 5

Results For Intended Task Development

	Training Program		
	Mechanic	Programmer	Secretary
No Development Intended	52	104	51
Some Development Intended	128	76	129
Level Of Task Development:			
Other Than Ability To Perform	7	3	8
Basic Ability To Do Task	20	28	44
Performance Ability Plus Standards Of Speed, Accuracy, And/Or Excellence:			
Level 5 (Increasing Levels Of Proficiency)	59	23	53
Level 6 (Increasing Levels Of Proficiency)	42	20	21
Level 7 (Increasing Levels Of Proficiency)	0	2	3

Of the 180 tasks listed in each survey, no development of 29% of the Mechanic tasks was intended, the same for 58% of the Programmer tasks and 28% of the Secretarial tasks. Comparatively little training of Programmer tasks was intended prior to employment, with less than half the tasks that were relevant to the job being selected for training. This small number of tasks receiving preemployment training is reflected in the correspondingly low number of training emphasis areas that were identified. By the way, this result corresponds to information we got from employers, who said they generally train programmers themselves, and do not expect much training prior to employment. Of the tasks for which some training was intended, meaningful performance standards existed for 80% of the Mechanic tasks, but for only 60% of the tasks trained in the other two types of programs. Training intentions can be described in abbreviated form for each task, as illustrated briefly in Table 6, using examples from the Secretarial occupation.

Table 6

Sample Task Training Intentions

Intention			
Level		Emphasis	Task
0	No Development		Assemble and Staple Dupli-Cated Materials.
3	Less Than Ability To Perform	Relating To Others	Greet Callers Or Visitors.
4	Develop Basic Ability To Perform (no special standards)	Basic Education	Place Telephone Calls.
5	Develop Ability To Perform With Advanced Proficiency		Proofread Type-written Copy.
6	Develop Ability To Perform With Advanced Proficiency	Technical Knowledge	Edit Letters Dictated By Employer.

For some of the readily learned tasks, such as the "assembling and stapling of duplicated materials," no development was intended in pre-employment training. Some other tasks did not warrant the development of performance ability, but the training did intend to incorporate an emphasis upon at least one task area other than performance of the task itself.

In the examples on this illustration, training in task performance was intended for the last three tasks listed; the first with no special standards of performance, but including an emphasis upon the learning of some elementary communication skills. Some rather advanced proficiency in task performance was intended for the last two tasks, including an area of special emphasis for one of them.

Uses For The Curriculum Content Questionnaire

The intention of portraying the curriculum content of a training program in this manner is to allow direct comparisons to be made with the results of task surveys and other front-end analyses resulting from the systems engineering and instructional design of training programs. Whatever the form in which the present curriculum content may exist, the use of the Curriculum Content Questionnaire appears useful in converting that actual content to a form more compatible with task survey data and analyses of the job. It indicates, if some training for a task is intended, to what general level of proficiency, and what areas of training emphasis are intended. It serves to identify what job capacities get developed, though not the enablers entering into such learning and development. These elements should permit a reasonable comparison to be made with the results of analyses identifying performance training needs of an occupation.

In addition to its application in this context, use of the Curriculum Content Questionnaire would also appear to be of potential utility in several other matters. For one, it could be used to develop a composite picture of a training program where that training occurs at different locations or through a series of instructional courses, such as might occur in local unit training. In another instance, it might serve as a useful means for denoting the intended skill level of trainees on particular tasks, for use in developing samples of job performance measures or other work sample tests, or for use in communicating the intention of the school to operational units (or other such employers of the graduates) so that local units might plan appropriate assignments and subsequent on-the-job training.

While this method of relating task surveys to the content of existing training programs may not represent a final satisfactory solution to the problem, it is hoped that it may perhaps serve your present needs and also stimulate the development of even more useful methods.

References

Ammerman, H.L., & Essex, D.W. Performance content for job training. Volume 4: Deriving performance requirements for training (R&D Series No. 124). Columbus: The Ohio State University, The Center for Vocational Education, March 1977.

Johnson, M., Jr. Definitions and models in curriculum theory. Educational Theory, 1967, 17, 127 - 140.

Johnson, M., Jr. On the meaning of curriculum design. Curriculum Theory Network, Spring 1969, 3, 3 - 9.

Walker, D.F., & Schaffarzick, J. Comparing curricula. Review Of Educational Research, 1974, 44 (1), 83-111.

Winer, B.J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

APPENDIX - DIRECTIONS FOR RATING CURRICULUM CONTENT INTENTIONS

INSTRUCTIONS FOR QUESTION I

Question I asks what level of task development is intended?

For each of the listed job tasks, CIRCLE THE APPROPRIATE NUMBER "0" to "7" TO INDICATE THE LEVEL OF TASK ABILITY YOUR PROGRAM INTENDS TO DEVELOP IN STUDENTS.

LEVELS OF TASK DEVELOPMENT

This is a rating of the extent that the curriculum, during the training program, deliberately plans to develop task proficiency. The following answer scale is to be used.

0 = NO DEVELOPMENT of the task is intended.
1 = Develop only a GENERAL AWARENESS of the task.
2
3
4 = Develop a BASIC ABILITY to perform the task.
5
6
7 = Develop a VERY HIGH PROFICIENCY in the skillful performance of the task.

Use all eight scale levels, as appropriate.

The ratings of "2" and "3" represent intermediate levels of ability between "GENERAL AWARENESS" and "BASIC ABILITY." Similarly, the ratings of "5" and "6" represent intermediate levels of ability between "BASIC ABILITY" and "VERY HIGH PROFICIENCY." Thus, the scale represents a series of increasingly higher levels of skill development prior to employment in the occupation. Level 4 implies no special intent that any advanced speed, accuracy, or excellence of task performance be developed.

Levels of task development should be those which are presently planned to be part of your curriculum, consistent with the learning abilities and employment opportunities of your students, and recognizing that much learning and skill development must come from actual on-the-job experience.

INSTRUCTIONS FOR QUESTION II

Question II asks what training issues, if any, do you intend to particularly emphasize in the training of a task?

Listed below are some suggested training issues that might be emphasized for a task. For convenience these task-related issues are grouped under three categories: JOB CONTEXT, PERSONAL, AND TECHNICAL. Please take the time to carefully review and understand each issue.

Where some feature other than performance ability is especially important in your training for a task, PLEASE ENTER A NUMBER from "1" to "11" TO INDICATE A PARTICULAR TYPE OF TRAINING ISSUE THAT IS EMPHASIZED. More than one issue may be entered for a task. However, an issue should not be marked for each task, but only where one or more issues is to be particularly emphasized in training.

WRITE IN ANY ADDITIONAL TRAINING ISSUE YOU INTEND TO EMPHASIZE.

SOME ISSUES THAT MIGHT BE EMPHASIZED IN TASK TRAINING

JOB CONTENT

1. TASK ORDER AND TIMING

   Recognizing when to do a task, particularly its proper sequence in relation to other work being done on the job.

2. TASK VALUE AND PURPOSE

   Sensitivity to the actual job value, usefulness, need, or importance of a task.

3. SAFETY

   Knowledge of the safety procedures and precautions that should be observed when doing a task.

4. VARIED WORK CONDITIONS

   Ability to do a task under a variety of conditions, events, or circumstances that may occur on the job. These may include hazardous, uncomfortable, uncertain, or stressful situations which place special demands upon the worker.

PERSONAL

5. RELATING TO OTHERS

   Developing skill in relating to other people at work. This may involve teamwork or cooperation, or personal skills in dealing with such people as customers, officials, other workers, or the general public.

6. WORKER ATTITUDE AND RESPONSIBILITY

   Developing special pride in work done. This may deal with feelings toward doing quality work and meeting performance standards, or it may involve personal work habits that influence how well a task gets done (such as being careful, or attentive to details).

TECHNICAL

7. BASIC EDUCATIONAL SKILLS

   Learning the particular elementary reading, writing, arithmetic, or speaking skills needed for effective performance of a task. (This category does not include advanced technical development of literacy and computational skills.)

8. DETECTING DISCREPANCIES

   Recognizing and interpreting the key events and conditions that indicate when something is not meeting performance standards or is deviating from acceptable tolerance limits.

9. TECHNICAL KNOWLEDGE

   Knowing and understanding certain key information, or a particular technical concept that has practical use in performing a task. This might involve knowledge of: vocabularly and nomenclature, subject-matter content, machine characteristics and specifications, organizational or system structure, advanced computational skills, operating principles and theories, rules and standards, or other such technical information.

10. SUPPORTIVE JOB AIDS

    Use of a job aid that assists the worker in doing a task more effectively or efficiently. These aids might be tools, charts, test instruments, checklists, reference guides, templates, procedural manuals, maps, forms, wiring diagrams, or other such devices and memory aids that support task performance.

11. ALTERNATIVE METHODS

    Capability for doing a task in more than one way or upon more than one type of thing acted upon. Such flexibility may be important for performing unusual or emergency procedures that may be required in actual job performance. In some instances flexible approaches are needed for work on different brands or types of equipment and material.

EXAMPLES:

Before beginning, please study the following examples of the rating procedure:

Tasks of the Job	Question I: Circle the Level of Ability Intended to be Developed in Training. 0 = No Development 1 = General Awareness 2 3 4 = Basic Ability 5 6 7 = Very High Proficiency	Question II: Where Some Feature Is Especially Important, Enter the Task Issue(s) to Be Emphasized. Job Context 1. Order/Timing 2. Value/Purpose 3. Safety 4. Varied Work Conditions	Personal 5. Relating to Others 6. Attitude & Responsibility	Technical 7. Basic Education 8. Detect Discrepancies 9. Technical Knowledge 10. Job Aids 11. Alternate Methods
1. Advise contractors on procedures, limitations, requirements, and delivery schedules.	0 1 2 3 4 5 6 7			
2. Organize guided tours for visiting groups.	0 1 2 3 4 5 6 7	2	5, 6	
3. Operate dictaphone.	0 1 2 3 4 5 6 7	Familiarity with more than one system of shorthand.		
4. Transcribe (type) from shorthand machine tape (stenograph).	0 1 2 3 4 5 6 7			
5. Run errands.	0 1 2 3 4 5 6 7	4		

In the first task, no training is provided, and therefore no issues can be emphasized.

For task two, basic ability to do it is intended, and it is important for students to learn the particular job value of the task. Special training of interpersonal skills is also emphasized, as well as a full sensitivity of the need to take particular care in organizing efficient and informative tours.

For task three, basic ability in task performance is again planned, but no task issues are especially emphasized.

For task four, some skill in performing accurately is intended, and ability to transcribe from more than one system of shorthand is emphasized in training.

For task five, little actual performance ability is being developed, but special attention is given to the various situations under which such errands often need to be done.

BEGIN YOUR ANSWERS BELOW
PLEASE RATE ALL TASKS ON QUESTION I BEFORE ANSWERING QUESTION II

CS

Applying Occupational Survey Data in Instructional Systems Development

Hendrick W. Ruck
Richard T. Dineen
Claude C. Cunningham

Although it is generally agreed that occupational survey data are extremely useful in developing training requirements, current Air Force procedures do not specifically govern the application of such data in instructional systems development. This paper presents a model developed by the authors which is based on current Air Force instructional systems development (ISD) directives. The model, therefore, is not empirically based, rather, it follows established principles of ISD. The model is proposed as a guideline for training decisions with the full expectation that it will be modified as necessary.

Existing procedures for developing training requirements for Air Force specialties normally result in Specialty Training Standards (STS) which are not task-specific. STSs are listings of items requiring formal training for airmen in a given specialty. Correlating occupational survey task statements with STS items is a difficult, laborious chore which may not always be performed due to the amount and difficulty of work involved. A starting point for alleviating this situation would be the changing of STS item statement to occupational survey task statements. While this model assumes that such a change will be made, the usefulness of the model would not be lessened should the change be rejected, since newly developed computerized matching of STS items and occupational task statements is expected to be available in the near future.

The basic model proposed here is comprised of three different decision subsystems. The subsystems flow sequentially, however, the second and third subsystems maybe applied concurrently. First, tasks are selected for training and placed on the specialty training standard. Second, task skill-knowledge codes are assigned. And third, the formal, basic school training course is derived. The most significant underlying assumption of the model is that occupational survey data are a valid measure which may largely determine training requirements.

In the interest of clarity, a definition is appropriate here. The work task has been used in this paper in referring to occupational survey data. Training experts often use the word task, and their definition of this word may differ from that meant with respect to occupational survey data. A task is defined here as a behavior which is time measurable, that has a beginning and an end, and that is understood to be performed in only one way. Performing a procedure, if the procedure is invariant, may be a task. Trivial actions, such as inserting keys or removing specific screws, are not usually considered to be tasks: rather they are elements of tasks, and, therefore, are not specified in occupational survey data.

Selection of Tasks for Training

Under this model, the specialty training standard is developed by listing all tasks performed by at least ten percent of the job incumbents of at least one of the skill levels (the Air Force Specialty Training Standard is designed for apprentice, journeyman, and technician training requirements). That is, tasks performed by ten percent or more of the job incumbents at any skill level represent a comprehensive and valid listing of tasks applicable for training in the specialty. Indication is made on the standard as to which tasks are applicable to the various skill levels, since not all tasks are relevant at all skills.

In developing the above criterion, initially four task factors were considered, as suggested in the Air Force literature. The factors were: percent members performing, learning difficulty, probable consequences of inadequate performance, and task delay tolerance. Using cutoffs derived from the literature, the criteria were used to develop training standards for three Air Force specialties. An analysis of the resulting training standards suggested that the ten percent cutoff on percent members performing was an adequate substitute for the complex four factor criteria suggested by the ISD literature.

Assignment of Task Skill-Knowledge Codes

Once tasks are selected for training, the establishment of skill-knowledge or proficiency codes is the next logical step. This step is not presently being recommended for implementation due to conceptual and research questions which arise in developing guidelines. The problems encountered in attempting to establish skill-knowledge codes is described to emphasize the need for further consideration of this topic by policy makers and researchers alike. The third subsystem of this model is not dependent on this step, therefore, the actually applicable model may be viewed as a two step model for the present.

It was at this juncture that difficulty in applying ISD guidance was encountered. Six criteria were selected based on ISD literature. They were: percent members performing, task difficulty, task delay tolerance, probable consequences of inadequate performance, frequency of performance, and time to initial performance. Measures of the first four criteria were derived from exact measurement (percent members performing was based on a count, the other criteria were based on ratings). Global estimates of frequency of performance were made based on cumulative time spent over all tasks. Finally, time to initial performance was considered to be "low" if 30 percent or more of the airmen within the first year following training performed the task, otherwise it was considered to be "high."

After developing complex decision rules based on the ISD literature, two problems arose. First, it became quite clear that decision rules involving all factors were not necessary for establishing proficiency codes. Percent members performing affected inclusion into the training standard, as did time to initial performance. Also, frequency of performance did not seem to affect the proficiency codes in any meaningful way. The only factors found to actually affect proficiency code levels were criticality (as measured by probable consequences of inadequate performance and task delay tolerance) and learning difficulty. More disturbing, however, was the conceptual problem of whether proficiency level is invariant for a task or whether, in fact, proficiency changes over time as individuals learn more about the task and gain additional experience. ISD literature is not clear on this question. The assumption that adding elements (skills and knowledges) to an occupational survey task changes the task such that it is a new task further confuses the issue. It was felt that until research findings and policy decisions shed light on these problems, the proposed model could not realistically address the issue.

Derivation of Formal Basic School Training

Once the tasks are listed on the training standard, percent members performing for first enlistment personnel is the primary factor for inclusion in the resident course. In addition, extremely difficult or extremely critical (as measured by probable consequences of inadequate performance and task delay tolerance) tasks are included in the course. Provision for manual override is made in the model, with the understanding that overrides will be justified. The algorithms used in the model can be programmed so that occupational data can be displayed and tasks flagged to facilitate the construction of the specialty training standard.

SUMMARY

This paper presents the results of a concerted effort to interface occupational survey data with instructional systems development. Unlike other ISD literature, this effort was data based, using data from three different Air Force specialties. The use of real world data in applying ISD criteria allowed for a reassessment of the criteria. Although the model has been applied to one Air Force specialty by training specialists, further testing of the model is required to provide a decision-data-base which will allow for adjustment of cutoffs and criteria.

It is important to note the difficulty encountered in developing proficiency codes. Certainly, further research is required on this point.

Finally, this paper oresents the beginnings of a model which was derived from existing models and actual data. The model offers a data-based development of courses as well as the capability to expand and contract courses objectively. The model is not finalized, it is not being seriously considered for adoption. However, with the appropriate testing and development, it offers the hope of objective proceduralized training development for all training systems which have access to occupational survey data.

Implementing Instructional Technology in Army Training: Some Obstacles and Solutions

Wesley K. Roberts
Courseware, Inc.
San Diego, California

This paper presents a summary of contractually supported instructional technology assistance provided to the various Army service schools through the U.S. Army Combat Arts Training Board during the period 1971-1975. Assuming the role of a change agent is challenging without the presence of obstacles; implementing modern instructional methods into Army training proved to be full of obstacles as well as a challenge. Only through a tenacious effort on the part of a few Army officers and noncommissioned officers to meet the change agent challenge have many programs prominent in Army training today survived.

The content of this paper is candid and is not available in any other singular source. It provides a historical perspective on many Army training programs today and the effort required to prepare the system for their implementation.

On 17 December 1971, the Army's Chief of Staff authorized the formation of the U.S. Army Combat Arms Training Board at Fort Benning, Georgia. This action was the result of a recommendation by the Board of Dynamic Training which had been functioning since August 1971 as an investigative agency to consider ways of supporting unit commanders in conducting meaningful and exciting training. Formation of the U.S. Army Combat Arms Training Board allowed for implementation of recommendations made by the Board for Dynamic Training.[1]

Identifying a Requirement for Instructional Technology Training Support

A major recommendation by the Board for Dynamic Training was that a program be developed that would rapidly restore deficit Noncommissioned Officer and Specialist confidence by upgrading training to develop their professional competence. To meet the challenge of correcting the gaps (cf. Kaufman, 1973)[2], identified in the training system, it would be necessary to systematically improve training provided to combat arms soldiers and ultimately to all soldiers. Some requirements were isolated for immediate implementation from which the generic effects of a systematic approach could be further implemented.

A priority requirement was for the combat arms schools to provide Military Occupational Specialty (MOS) related training extension courses directly to soldiers in small units. These courses were to be prepared using a multimedia format, directed at both individual and small group training. The concept of the training extension courses was to take subject matter expertise found in the Army's service schools and export it to the soldiers "in-the-field" in the form of up-to-date training materials. Through this method, individual training in units would be kept current with service school doctrine.

In order to begin this extensive program of extension course development, it was necessary to begin to identify and make provision for the support required. A cursory survey of the combat arms schools (Air Defense, Armor, Field Artillery, and Infantry), and feedback from the implementation of similar programs, indicated a need to provide training for this requirement not only at the technical personnel level, but for middle and senior managers as well. Crucial to the mission of preparing these extension courses was support at the command level within each combat arms school.

To facilitate the senior level management support for the extension course program, a conference for Assistant Commandants of the combat arms schools and representatives from the Department of the Army was held at the United States Military Academy at West Point, New York. During this conference, attendees were briefed on current techniques in instructional technology, as well as the major obstacles to be removed to partially insure success of the program.

As the plan for the development of the training packages in extension course form, ergo, Training Extension Course (TEC), was taking shape; considerable effort began to identify additional training support needs. A review of the existing training regulations and guidance (CONARC, later revised and adopted as TRADOC Regulation 350-100-1 and the then existing FM 21-6) provided the following insight as to the nature of what training may be required. Deficiencies identified in the CONARC Reg 350-100-1 included:

(1) little "how to" guidance for training developers,

(2) no overview of the total system,

(3) criterion tests were developed after training materials,

(4) knowledges and skills are not related methods,

(5) no developmental model,

(6) nothing addressed the actual conduct of training, and

(7) it contained only a cursory section on quality control.

Another source of input for training support requirement was FM 21-6, Conduct of Military Training. At that time, FM 21-6 addressed resident instruction and did not provide guidance concerning other training methods.

The training support requirement expanded when job task data information at the service schools, as specified in 350-100-1, was found to be insufficiently developed. Further, the Military Occupational Data Bank (MODB) did not have the required job analysis information for designing required instruction. At that time, the information available was aggregated for personnel purposes. Due to the deficiencies identified, it became necessary to initiate several research and development programs.

A contract with the Human Research Resources Organization (HumRRO) was established for the purpose of analyzing job requirements for the eight Military Occupational Specialties selected for initial TEC development (11B, 11C, 11D, 11E, 13A/B, 13E, 16P, 16R). This action was taken to resolve the immediate job analysis data needs for systematically developing training in the TEC program and to serve as an empirical basis from which new directions in occupational research, in support of training requirements, could evolve.

A systems engineering workshop was held to instruct service school personnel how to generate front-end analysis data; how to write task statements, prepare task lists, surveys, etc. This workshop was held at several sites, conducted jointly by HumRRO and Combat Arms Training Board personnel, and attended by service school personnel identified to work in the TEC program.

Further, professional support was provided to the emerging TEC program by contractual assistance provided through the Army Research Office and Battelle Laboratory's Durham, North Carolina Office. This allowed the Combat Arms Training Board to acquire the expertise of several professional analysts who were knowledgeable in the field of Instructional Technology. These analysts contributed technical guidance to managers in TEC program and other emerging programs. This service became a valuable asset in the TEC program and tremendous benefit resulted from its use.

More Evidence for Training Support

Concurrent with the conceptualization of the design to be used in developing TEC lessons, planning was made to test this approach by providing training to soldiers preparing for the 1972 11B40 (light weapons infantryman) MOS Test. This phase, known as the Unit Training Extension Course (UTEC) Program or TEC I, was used to test the effectiveness of the TEC concept. Scores on the November 1972 11B40 MOS Test were used as the criterion of effectiveness.

A total of 56 TEC lessons were developed by a committee within the U.S. Army Infantry School, addressing four major test domains. These lessons were largely prepared by Infantry Officer Advance Course students and were in a 35mm slide/synchronized sound cassette tape format. HumRRO Division No. 4, Fort Benning, Georgia was contracted to design a study to evaluate this project.[3]

Results of this investigation reported positive gains from the use of these instructional materials by soldiers in preparation for the 11B40 MOS Test when verbal ability, command emphasis, and study factors were weighted with test scores. Otherwise, no significant differences were found between soldiers using and those not using the instruction.

A deficiency found in the conduct of this early TEC effort were the materials themselves. The lessons were not specifically designed to teach those skills being evaluated on the 11B40 MOS Test. The relationship between the TEC I lessons and the test items evaluated by the November 1972 version of the 11B40 MOS test was ambiguous with the exception of one major test domain. These findings are not surprising in retrospect: the developers of these lessons had no training in instructional technology prior to their involvement in this project; the lessons themselves were simply "illustrated lectures", even though they represented a noteworthy effort; objectives had not been systematically prepared from use of a job analysis; and the 11B40 MOS test items did not measure mastery of TEC I lesson objectives. Further, systematic revision of lessons that did not adequately instruct their objectives was not provided for during the developmental process. A recommendation made by HumRRO, as a result of this study, was to "systems-engineer" all future TEC lessons, allowing for the lessons to concentrate on job-relevant skills and to be evaluated by job-relevant performance tests.*

The Requirement for Instructional Technology Training Support

Abundant empirical evidence clearly indicated that the Army's combat arms schools lacked personnel trained to prepare the systematic instruction demanded if the TEC program was to achieve the goal of rapidly restoring Noncommissioned Officer and Specialist confidence and competence (in part) via extension courses as recommended by the Board for Dynamic Training.

*During this time frame the movement for incorporating performance tests into the MOS testing program was gaining considerable momentum but was not conceived in the present Skill Qualification Test form. Performance testing was largely restricted in use to Basic Combat Training. The results of this and similar tests aided in the move to a more logical method of job performance evaluation being implemented under the current Enlisted Personnel Management System.

Existing instructor development programs were evaluated as not adequate to train personnel in instructional technology. The Army's systems-engineering model (them CONARC, now TRADOC Regulation 350-100-1) had been reviewed[4] and was found to be an inadequate source of instructional development guidance.

A search for an existing training course in Instructional Technology outside the Army was then conducted. During this search a lesson development workshop that could immediately facilitate training of personnel for the TEC program was identified. The model taught in this workshop is known as CISTRAIN, the acronym for Coordinated Instructional Systems.[5]

Two Instructional Technology Workshops (CISTRAIN) were contracted by the Combat Arms Training Board to Deterline Associates, Inc. The first of workshops was held in San Francisco, California. Attending this 28 November - 8 December 1972 workshop were 35 personnel from the Army's Combat Arms Schools, the Combat Arms Training Board, and the United States Military Academy and other agencies. The goal of this workshop was "to enable workshop attendees to further develop and strengthen their knowledge concerning instructional technology for the purpose of conducting similar training for staff and faculty members of their respective schools that would be directly involved in the development of TEC II instructional materials ... and to further develop staff and faculty not only for designing and producing TEC II instructional materials, but also to expand instructional technology expertise throughout the Army school systems." [6]

Concurrent with this training, specifically in April of 1973, Robert K. Branson and Robert Morgan of Florida State University conducted a two day seminar for assistant commandants of the Army's Combat Arms Schools and other high level officers. The purpose was to inform high-ranking officers of the implications of instructional technology for large scale planning, critical to the success of the long-range effort.

The second Instructional Technology Workshop (CISTRAIN) held in support of the TEC program was conducted during the period 9 - 20 July 1973 in Washington, D.C. At this workshop, 37 attendees from Army Service Schools, the U.S. Army Combat Arms Training Board, the Air Force, the Navy, and the Marine Corps received instruction similar to those attending the earlier San Francisco Workshop. The goals for the second workshop were somewhat modified from those in the earlier one. They were: "(1) to broaden the U.S. Army's community of education technologists by offering training by some of the nation's top educators, (2) to further develop and strengthen the knowledge and skill of workshop participants so they can conduct a similar course for staff and faculty members of their respective schools and, (3) to provide the participants with an instructional development model and teaching vehicle that can be used to train instructional developers in U.S. Army Service Schools who develop training materials for resident courses or for export to soldiers in units."[7]

The Washington, D.C., workshop represented the training identified as a necessary catalyst for follow-up instruction at five service schools who were to become the participants in the TEC program now known as TEC III. These were the U.S. Army Engineer School, the U.S. Army Southeastern Signal School (now the U.S. Army Signal School), the U.S. Army Quartermaster School, the U.S. Army Ordnance Center School and the U.S. Army Adjutant General School. Each service school in attendance was provided with a complete set of workshop materials to meet the requirement of further training for the TEC program personnel.

Although considerable technical support and guidance was being provided to Army service school personnel for the TEC program, it was necessary to expand training support for the long-range instructional system development that was to be ultimately implemented.

The United States Army Infantry School, supported by Combat Arms Training Board funding and contractually assisted by Insgroup, Inc., began an extensive evaluation of its instructor training course. As a result, it was determined that additional instruction was required on the theory of learning and lesson development. All instruction was converted to self-paced, mediated materials, i.e., TV, tape-slide programs and programmed texts. The number of practical exercises was reduced from nine to seven by eliminating one twenty-minute exercise and a briefing requirement. The revised self-paced course was an improvement, but still did not account for diversified and individual training requirements.[8]

The Fading of In-House CISTRAIN Model

Instructional Technology Workshops

At the conclusion of both the San Francisco and Washington, D.C. workshops, it was intended that the participating service schools would use the workshop materials provided to conduct subsequent workshops to train a substantial number of people in the basics of instructional technology. Nevertheless, this program was never fully realized. Although several schools made noteworthy attempts, only the U.S. Army Armor School (who sent their Education Advisor to the course) ever really implemented follow-on training using these materials. In retrospect, some reasons for this failure may be intuitively identified. CISTRAIN was not Army doctrine nor was it ever officially sanctioned as a faculty development program to be institutionalized in the Army. Service schools did not (in many cases) send a faculty developer, i.e., an instructor for instructors, to the workshop - hence who was going to subsequently follow through with the training program? Many high level managers did not perceive this to be a viable solution to their training problem (instructional technology was new to them) and did not choose to support the training of additional personnel. There was no follow-on training provided to potential course managers to insure they could teach the workshop. And, the schools perceived (in many cases) that a contractually supported effort for the development of TEC lessons relieved the necessity for such training.

The Broadening of the Training Requirement

During the period when the TEC program was contractually adding "trained" Instructional Technologists to the organizations participating in this endeavor, the U.S. Army Combat Arms Training Board was busy supplementing this training with internally developed seminars, conferences, and workshops designed for various tasks necessary in developing systematic instruction. The zietgeist prevailing in the Army training community was emerging as that of meeting the training requirements head-on with the latest developments in instructional technology. Said another way, the "tip of the iceberg" which would lead to the total commitment in the Army toward the full exploitation of training technology during the middle and late 1970's was beginning to emerge.

Contemporary issues at the national level were often leveled at the post-Vietnam military establishment in terms of substantial budget cuts. Congressional leaders were being forced to call for "more bang for the buck" and "more teeth than tail." Consciousness was being raised at the highest levels across the armed services of the desirability of less time spent in training and more time in deployable combat and combat support units; toward higher levels of professionalism on the part of armed service personnel; toward increased efficiency and proficiency in job performance; and all of this, for less training cost.

One alternative to the tremendous pressures of the day was to conduct interservice training when the tasks to be trained were common across one or more service. While there has always been an exchange of training ideas and some programs throughout the history of our armed forces, it has not been without trepidation of the "Purple Suit," wherein the various military services lose their distinct identity, a concept insensitive to the ever prevailing philosophy that our specialized armed services (e.g., land, sea, air) require specialized training to perform their distinctly assigned missions. Additionally, there are the manpower and budget consequences of turning over a portion of training to another service. These considerations provide understandable reasons for the military establishment's reluctance to conduct large-scale interservice training.

In an effort to ferret out solutions to the dilemma posed by the training resource and manpower constraints being placed on the post-Vietnam armed services, the commanders of the four military commands met in Washington, D.C. during the September of [illegible] to establish an Interservice Training Review Board. The purpose of the Board was to promote economy in training through the use of interservice training. Subordinate to the Board were several committees constituting the Interservice Training Review Organization. Each of these committees were further broken down for more specific functions for which subcommittees were formed. One of these subcommittees, the Interservice Subcommittee for Instructional Systems Design, was formed when its predecessors function (standardization of a training glossary for the armed services) was redirected. Concomitant with this redirection was the broadening of the new subcommittee's function. It was charged to develop a model and set of procedures for the development of curriculum for interservice training programs.[9]

Paralleling these activities were several on-going projects within the Army independently working toward partial solutions to the larger issue of systematic training development and management for the total training system. There was the HumRRO 8-MOS study, the revision of TRADOC Regulation 350-100-1, research funded by the Army Research Institute on Criterion-Referenced Tests, the TEC program, a Baseline study designed to identify common, semi-common, and unique soldiering tasks, the Experimental Volunteer Army Training Program (EVATP) at Ford Ord, California, the need to assist service schools in expanding instructional technology skills to more personnel, and the potential adaptation by the Army of Air Force Pamphlets 50-2 and 50-58 on Instructional Systems Design. In order to capitalize on this and other research and development efforts, the Combat Arms Training Board formalized a contract with Florida State University for the purpose of assisting the Army in bringing it's spectrum of training research and development into a total working system.

The Combat Arms Training Board and the Interservice Project

Not oblivious to, yet separate from the Interservice action mounting in training circles in the early 1970's, the Combat Arms Training Board, in fulfilling its charge stemming from the recommendations of the Board for Dynamic Training, and in an effort to investigate the actual status of instructional capabilities within the Army were, had, on 29 May 1973 entered into a contractual arrangement with the Center for Educational Technology of the Florida State University. The original purpose of this contract was to survey the state of the art in Instructional Technology in the Army.

Preliminary reports from the Florida State Study cogently indicated a requirement for the Army to develop an instructional technology manual which would lend harmony to the guidance and initiatives in training research and development. Based on these early findings and recommendations, the Combat Arms Training Board expanded the scope of work under contract with the Florida State University to allow for the design of an instructional technology manual and supporting workshop materials.

On 6 September 1973, members of the U.S. Army Combat Arms Training Board met in Atlanta, Georgia with Major General Ira Hunt, the Deputy Chief of Staff for Individual Training, U.S. Army Continental Army Command (now reorganized and the Service School and Training Center mission is assigned to the new U.S. Army Training and Doctrine Command - TRADOC). General Hunt was briefed on the findings and recommendations of the study performed by the Florida State University. Based upon this report, General Hunt approved the Combat Arms Training Board to continue its contractual development of an instructional technology manual which would be designed to replace CONARC Regulation 350-100-1, and the necessary workshops designed to train and implement the new instructional technology manual at the service school top level management, middle management, and instructor level.[10]

Earlier, on 26 July 1973, the Interservice Subcommittee on Instructional Systems Design met at Fort Benning, Georgia, to resolve procedural questions in connection with the design and development of a model and set of procedures for Interservice curriculum development.[11] This meeting was called in light of the committee's knowledge of the contract between the U.S. Army Combat Arms Training Board and the Florida State University. This meeting led to an agreement between the U.S. Army Combat Arms Training Board, the Florida State University, and the Interservice Subcommittee on Instructional System Design, that the contracted work could possibly be redirected without serious imposition to serve the needs of not only the Army, but the interservice community as well. This meeting led to the eventual inclusion of Interservice participation in the Combat Arms Training Board's contract with Florida State University.

While the Interservice Committee had the charge of developing a model and set of procedures for interservice curriculum development, the U.S. Army Combat Arms Training Board required broader research and development investigation for the improvement of Army training. Therefore, from time to time, modifications were made and tasks added to the contract with the Florida State University that addressed the specific requirements of both the U.S. Army Combat Arms Training Board and the Interservice Committee. The research and development products delivered by the Florida State University ranged from a report summarizing the state of the art in instructional technology in the Army, a prototypic TEC audiovisual kit, a manual for preparing extension courses, a model for service school staffing, and a model and set of procedures for interservice curriculum development.

The Interservice Procedures for Instructional Systems Development

The singularly most significant product developed by the Florida State University, the one most often referred to, was the five-volume set of manuals titled Interservice Procedures for Instructional Systems Development.[12] These procedures are often thought of as the only product resulting from the Combat Arms Training Board - Florida State University contract. This is not true. Although their preparation could not have been accomplished without some of the products from other contractual tasks, these manuals represent part of the deliverables from Task 5 only (there were a total of eight contractual tasks). It is important to note here that the major intent of Florida State University's research became that of preparing a manual for TRADOC schools. It was subsequently modified to the preparation of a set of manuals on Instructional Systems Development for Interservice Training.

Input to and the development of these procedures was a massive process. Possibly the most difficult task performed by the authors of the procedures was to restrict the narrative to content pertinent to interservice curriculum developers. One of the methods used to insure the content of the procedures was relevant to their intended target population was to conduct formative evaluation trials on both the procedures themselves, and the three levels of training workshops designed to support their implementation and use.

Formal review of the procedures began during November 1974. Phases I and II (Analyze and Design) of the procedures were evaluated in an interservice workshop held at Fort Benning, Georgia. Phases III, IV, and V were evaluated in an interservice workshop held at the Naval Training Center, San Diego, California during February 1975. Further, the manuals were reviewed by experts in all Army Service Schools during the spring of 1975. The aggregated data and comments from attendees were used for revision of the procedures, and a revised version of the procedures was published in July 1975. Critical to the understanding of this program is it's separation from any particular program within any one service. The formative evaluation workshops were never intended to serve as catalysts for the promulgation of these procedures within the services. They were, simply, "validation trials." These workshops, for training the procedures to the technical, middle manager, and senior manager levels, were validated in workshops at Fort Benning, Georgia; Fort Eustis, Virginia; Naval Training Center, San Diego, California; Fort Gordon, Georgia; Tallahassee, Florida, and Pensacola, Florida.

Status of the Interservice Procedures for Instructional Systems Development

Although the model had been approved at the close of 1975, the revised version (July 1975) of the Interservice Procedures Manual was not adopted by the Interservice Training Review Board. The Interservice Committee for Instructional Systems Development prepared a report for the Interservice Board on pilot projects, implementing the procedures in actual training setting (c.f., Scanland, 1977).[13] The philosophy of the Board being one of insuring the procedures are reliable and usable, several pilot projects representing different research methods are planned by the services to evaluate the procedures. The largest effort was a contractually supported project at the U.S. Army Signal School, Fort Gordon, Georgia. This pilot implementation of the procedures did not serve the exact purpose of the Interservice Community. It was designed with the Army in mind under the philosophy that the interservice procedures were generic to all services, and specific enough to test within the Army. This project was never fully completed and was redirected to address job performance aids and other training methods. The Navy did not implement a pilot program; nevertheless, the procedures were published by the Navy as directive NAVEDTRA 106A. The Army published the manuals as TRADOC Pamphlet 350-30. The Air Force and Marine Corps did not assign an identifying number within their service to the interservice procedures. The general precept throughout the military training establishment became that of while the procedures were developed for an interservice purpose, they provided considerable "how to" guidance and reference for use a a resource document. Even though there was a continuing request for instructional technology workshops, these particular materials remained at the close of 1975 as relatively untested. The workshop materials and the procedural manual are currently being made available to the trainers within the various services, and it is anticipated that there will be several alternative methods employed to use these materials as training documents.

Future Requirements for Contractually Supported Generic Instructional Technology Workshops

Over the past several years (post-Vietnam era) the Army has been streamlining training and personnel management systems for a myriad of reasons. During 1974-75, two significant programs were developing which will have considerable and widespread impact in the Army over the next several years. These are the Enlisted Personnel Management System and the Officer Personnel Management System. The U.S. Army Training and Doctrine Command has participated in their development since the beginning of these systems. Calling for a significant restructuring of the training support for career development in both the enlisted and officer ranks, these systems provide for the multilevel structuring of training to support actual and potential personnel assignments. Concurrent with the development of these systems, considerable training resource and manpower constraints, the creation of three new combat divisions (for a total of 16 in the Active Army) within existing personnel limits placed on the Army, and other major considerations, have caused the U.S. Army Training and Doctrine Command to investigate other advanced technologies to achieve mission success within the tumultuous environment it operates in.

A very definite direction toward the self-pacing of instruction resulted as an outcome of an Instructional Technology Symposium held at Fort Eustis, Virginia in 1975.[14] The symposium focused on problems confronting resident Army service school training and the requirement to provide training to soldiers in units. The symposium resulted in two major directions to Army service school commandants. They were: (1) a charge from General William E. DePuy, the Commanding General of the U.S. Army Training and Doctrine Command at that time, to provide command level support to self-pacing initiatives, and (2) to use the self-pacing method whenever possible. This direction was directed at developing revised and new training programs in a more cost-effective manner and to enable soldiers to master resident school training based on their individual capabilities as measured against established criteria. The later effect was specifically an attempt to place soldiers into units as soon as possible by reducing their time in the training base.

Empirically, self-paced programs tend to reduce training time upwards of 25 percent and cut training costs overtime without loss in student performance. In many instances, the actual student performance has resulted in significant improvement over scores recorded in more traditionally conducted courses.

With the advent of the Enlisted Personnel Management System came the use of constrained task lists. This process was one of eliminating tasks from training programs by analyzing their relative importance to the overall goal of the course in which they were contained. The constraining of tasks to be taught resulted in reduction of some training time and consequently training costs. This procedure was by itself a major change in the direction from traditional training methods.

The development of Soldier's Manuals and their corresponding Skill Qualification Tests established Skill Qualification Tests established not only defined tasks for job performance within a Military Occupational Specialty, but the precise criteria for measuring task mastery. These two components of the Enlisted Personnel Management System are now the focal point for establishing remedial training for soldiers not performing at the required criterion level within their job on a task basis.

A Requirement for Assistance in Self-Pacing

As a result of many events culminating in the 10-11 December 1975 TRADOC Commanders Conference, a new requirement was identified to provide self-pacing workshops for Army training personnel. This requirement came about due to the accelerated nature of training technology developments within the Army to achieve its training mission vis-a-vis the economic pressures and manpower restrictions of the day.

In order to provide adequate training in the technology of self-paced instruction, the Training Management Institute (now the Training Development Institute) integrated relevant instructional materials that exist within the Army with pertinent, and essential, validated self-paced training materials and expertise available in the academic and industrial communities.

Contractual Assistance Required Beyond the Immediate Future

It would be a sophomoric assumption to state that the contractual support in instructional technology workshops completed in the recent years by funding support from the U.S. Army Combat Arms Training Board represents more than a beginning to what may be demanded in the long range future of training developments within the Army. Striving for excellence in training programs at such a rapid pace and with such tremendous in-house personnel and resource restrictions makes it inconceivable to identify, at this juncture, a training workshop in instructional technology either within or outside of the military which would be generic, yet, specific enough to address the training needs of the instructional technology personnel involved in improving Army training now and in the future. Therefore, the Army is preparing to offer a basic course in instructional technology with satellite workshops designed for more precisely defined training development and management tasks. The ability of the Training and Doctrine Command to meet its training missions of the future may be largely determined by its ability to provide rapid response to urgent training requirements with surgical precision.

Summary

To summarize the challenge that faced the Combat Arms Training Board in fulfillment of its change agent role for Army training, several obstacles may be identified. They are: (1) the existing information system was not prepared to address training problems; (2) there was a shortage of instructional technologists who could be instrumental in developing improved training materials; (3) responsible commanders required briefings, and in some cases convincing, on the benefits of implementing modern instructional methods, and (4) the total effort required centralized control and guidance to insure proper system integration.

The solutions to these major obstacles came about through (1) a large-scale effort to place job analysis information into the existing data base; (2) the conduct of workshops to provide initial instructional technology training to personnel involved in this effort; (3) briefings to inform responsible commanders and their staff of the immediate and long-range benefits of instructional technology to insure their support in this effort; and (4) the Combat Arms Training Board being established by the Army as the principal agency for control and guidance of implementing modern instructional technology into Army training.

A Final Note

This chronology has attempted to provide a historical perspective on major initiatives in instructional technology that were supported by the Combat Arms Training Board. However, many significant initiatives, i.e. training simulaltion, sub-caliber training devices and gaming simulaltion and others; were not addressed in this paper. This is not due to oversight, rather to the focus of the paper being on generic instructional technology training support and not on specific training programs.

References

1. Report of the Board for Dynamic Training, Volume II - Final Report. Fort Benning, Georgia: U.S. Army Field Printing Plant, 1971.

2. Kaufman, Roger A. Educational System Planning.

3. Jacobs, T. O. and Richard A. Hardy Jr., Test and Evaluation of Training Extension Course (TEC), HumRRO Technical Report 74-16. Alexandria, Virginia: Human Resources Research Organization, 1974.

4. Ricketson, Darwin S., RusselE. Shulz and Rogert H. Wright, Review of the CONARC Systems Engineering of Training Program and Its Implementation at the U.S. Army Aviation School: Consulting Report. Fort Rucker, Alabama: Human Resources Research Organization, Division No. 6 (Aviation), 1970.

5. Deterline, William A., and Peter D. Lenn, Coordinated Instructional Systems. Palo Alto, California: Sound Education, Inc., 1972.

6. TEC II Instructional Technology Workshop 28 November - 8 December 1972, San Francisco, California, Course Outline, Fort Benning, Georgia: U.S. Army Combat Arms Training Board, 1972.

7. Instructional Technology Workshop 9-20 July 1973, Washington D.C., Course Outline. Fort Benning, Georgia: U.S. Army Combat Arms Training Board, 1973.

8. Rutherford, Billy E. An Overview of the Trainer Development Program (TRADEP) at the United States Army Infantry School. Orlando, Florida: Naval Training Equipment Center Training Analysis Effectiveness Group Conference, 1974.

9. Branson, Robert K., et. al., Interservice Procedures for Instructional Systems Development: Executive Summary and Model. Tallahassee, Florida: Florida State University, 1975.

10. Memorandum for Record, Subject: MG Hunt Briefing on Florida State University (FSU) Instructional Technology Contract with USACATB, dated 11 September 1973, ATTS-TB-I, MAJ Thomas F. Plummer, Fort Benning, Georgia. USACATB 1973.

11. Memorandum to the Interservice Training Review Board, 30 July 1973 from CNET Code N-352 Chairman, Interservice Subcommittee on Instructional Systems Design, Subject: Minutes of Meeting 26 July 1973 at FortBenning, Georgia. Pensacola, Florida: CNET, 1973.

12. Branson, Robert K., Gail T. Rayer, John P. Furman, and J. Lamarr Cox, Interservice Procedures for Instructional Systems Development; 5 Volumes, Fort Benning, Georgia: Federal Printing Plant, 1975.

13. Scanlard, F. Worth. Final Report to the Interservice Training Review Organization on Instructional Systems Development. Pensacola, Florida: Chief of Naval Education and Training, January, 1977.

14. Report on Instructional Technology Symposium, Fort Monroe, Virginia: Headquarters, U.S. Army Training and Doctrine Command, 1975.

TRAINING EFFECTIVENESS MEASUREMENT:

ITS IMPACT ON PRODUCTIVITY

BY

CDR CARLTON F. MEREDITH, USCG
Chief, Training Division

and

MS. JOAN T. CHIPPENDALE
Education Specialist

USCG Training Center
Governors Island
New York, NY 10004

Military Testing Association Conference

17-21 October 1977

San Antonio, TX

INTRODUCTION

The U. S. Coast Guard Training Center at Governors Island is responsible for resident training in four general categories:

(1) Technical training to qualify personnel for job entry in seven specialty ratings;

(2) Advanced training in maintenance and operation of specialized equipment in four specialty ratings;

(3) Mission-related training for officers and rated petty officers in two mission areas;

(4) Program-related training for rated petty officers in two program areas.

For convenience, these categories are placed in two broad areas. Technical training courses for job-entry qualification are grouped into Class "A" Schools, the others are grouped into Class "C" Schools. Over 2000 students are graduated annually, approximately 50% in each broad area. There is great diversity among the individual categories, and even greater variety when these categories are broken into specific courses. For example, in the Electronics Technician (ET) job entry category alone, there is one basic core curriculum, and three advanced tracks (an integrated series of courses leading to qualification) which combined comprise a total of 36 individual courses.

Managing a Training Division characterized by so many unique Branches, Sections, Schools and Tracks is a significant challenge.

This is not a research paper, nor is it a "How to Do It" paper. It is intended to raise some questions, present problems which we face in common, and to seek assistance. The basic question is - How do we know that what we are doing is effective and productive?

As training managers, we need to have some way to measure the effectiveness and productivity of our training systems. These measurements may be as specific or as general as meets our needs, but in any case, we must be able to determine how well we are doing with what we have to work with. A military training manager's primary job is to meet the training needs of the "field". We must be concerned not only with the quantity of graduates, but with the quality of the graduate's performance as well. Secondarily, but closely behind in importance, we must insure that we are using the taxpayer's money wisely.

We have a distinct disadvantage in contrast to our counterparts in civilian education and industry. Most of us are not professional trainers. We are professional soldiers, sailors, or airmen who, by choice or otherwise, find ourselves in the role of training manager - a role which we may play for only a relatively brief period in our military

careers. Some of us may have had the advantage of formal training to prepare us for this role. For the majority of us, that is probably not the case. We have learned what we know about managing training mostly through the legacy of our predecessors. Ours is certainly not the ideal system, but it is one which has certain advantages when viewed in the overall military context. If our predecessors have built a strong and well-documented program, then we have something to build upon. If not, then we have to scratch for our own solutions until we can build our own functional management systems.

Our predecessors at Coast Guard Training Center Governors Island have been kindly. Nonetheless, we incumbents have found that no tools have been left behind to help us to make the vital training effectiveness and training productivity measurements. How do we know if our programs are effective? How do we know if they are productive? How do we know if changes are necessary? And if so, what changes? Where are our benchmarks?

BACKGROUND

The training philosophy at Coast Guard Training Center Governors Island is undergoing an evolution. Perhaps not as dramatically as the evolution of training in our counterpart DOD organizations, but an evolution nonetheless. We are progressing in all of the individual and integrated courses offered from traditional subject-matter based instruction to performance-based instruction. This evolution began in 1973 with the conversion of ET training. And, following our experience in that program, it is being carried out in all of the other rating areas as well. Because the conversion in ET training is complete, we will be using that as a primary example in this paper.

The previous ET graduate was a generalized specialist. He* knew a great deal about a lot of things related to electronics. He was well educated in his discipline, but not necessarily well trained to perform the technical tasks expected of him upon job entry at his first duty tour The ET could expect to be assigned to either a Coast Guard cutter or to an isolated LORAN station, and since they all received identical training heavily loaded with electronics "theory", none could be considered by present standards to be qualified to maintain and repair the specific equipment found at his first duty station. It was as if we were graduating apprentice electronics engineers rather than apprentice electronics technicians. Consequently, a period of OJT following job entry was required to enable them to do what their training should have prepared them to do. This was certainly not an effective nor a productive training method.

The performance-based system which we have implemented is designed to insure that the student can perform those tasks which will be required

*For convenience only, the masculine pronoun will be used throughout, rather than he/she or him/her.

of him upon job entry, no more and no less, and to perform them up to the requisite standards prior to graduation. His performance is reinforced through a series of "real life" performance tests in which he is required to trouble-shoot and repair the exact same type of equipment that he will be working with at his first duty station. So, by matching our training objectives closely to the actual job tasks, and our performance testing standards closely to job performance standards, we should be able to insure that the graduate is able to directly transfer his training experience to his work immediately upon job entry with minimal OJT.

We completed the ET conversion in 1976. We were satisfied that the new system was effective. After all, we were able to reduce the basic ET core curriculum from 20 to 12 weeks by cutting out all of the "nice to know" theory. Granted, the new performance-based training was an order of magnitude more expensive than the old subject-matter based system because of the necessity to procure more of the actual electronic devices used in the field, and more test equipment, but we were turning out better technicians. But, were we?

The first reaction from the field was negative. The new technicians "didn't know anything". No specific criticism, they were just "not as good as they used to be". These criticisms caused us great concern. We sensed that the old system was not effective, and was unproductive, but the field had been satisfied for years! We believed that the new system was extremely effective and productive, yet the field griped! The only way to insure that what we were doing was not only effective, but productive, and to respond to the field criticism, was to measure our effectiveness first, and then, using that data to see how productive we were. Direct and objective feedback on the performance of our graduates from the field was needed.

The decision was to conduct surveys of the performance of graduates who had been on the job for a minimum of 6 months and a maximum of one year. In order to be useful in measuring the effectiveness of training, the surveys had to be specific. Each job-task related to the newly learned technical skill had to be probed.

MEASURING EFFECTIVENESS

In order to base surveys of training effectiveness on fundamentals which would remain relatively constant, thereby enabling re-surveys to measure the same aspects for comparison, the supposition was put forth to formulate survey questionnaires on the Enlisted Qualifications Manual (CG-311), the paradigm of qualifying standards for each rate, to produce the quality product required. Conferences were held with the School Chiefs and Instructors to verify that CG-311, their curricula, and their Performance Objectives were indeed the same. Since Governors Island Training Center had recently reviewed and updated all Performance Objectives to meet the requirements of the Qualifications Manual, accordance was forthcoming, and the criterion was selected.

FORMAT

The format chosen for the questionnaires was a job task inventory developed from the tasks enumerated in CG-311 for E-4s. The imperative duties to be evaluated by the field were extrapolated, and final selections were made by a concurrence of Branch Chiefs and Instructors of the particular schools.

The questionnaire packet consists of a preparatory letter sent to the recent graduate two weeks before the arrival of the questionnaire. Two weeks later, the entire packet is sent to the graduate's Commanding Officer. It is requested that the Commanding Officer forward the graduate's questionnaire to the individual, that the Supervisor's questionnaire be forwarded to the person most qualified to evaluate the graduate's performance, and finally that the Commanding Officer complete a critique sheet evaluating the Supervisor's appraisal. An open-ended question is included in each packet for respondents to indicate "missing training elements", for possible future curriculum inclusion. Additionally, a reminder letter is sent to the graduate in the event his questionnaire is not returned within a month from the time of mailing. Both graduate's and Supervisor's questionnaires are comprised of an instruction sheet, a biographical data sheet, and the necessary number of pages of job task statements.

The utilization of simply-structured job task statements as the basis of the questionnaire eliminates a degree of generality by pinpointing specific learning elements. The respondents are asked to rate the job tasks in two catagories: Frequency of Performance of specific tasks in the field and Adequacy of School Training for the task. The response to Frequency is clearly objective; some element of subjectivity is apparent in the Adequacy rating, however since it is evaluation (opinion) which is sought, this is vital.

The number of tasks, categorized in major AREAS of performance, runs high. In all surveys, tasks averaged 105. This great number of tasks is deceptive however, for each one is composed of a simple statement, containing only one duty; one thought. A judgment is concise and succinct because there is only one issue to be evaluated. Since there is no delineating selectivity involved in the mental process, the time involved to weigh each task is short.

FEEDBACK METHOD

This feedback method provides a comparative study between choices selected by graduates and their Supervisors. It provides information as to (1) how the graduate views his job based on task performance frequency; (2) how adequately he feels he has been trained to perform each task; (3) how his Supervisor views the man's job task performance frequency, and (4) how well the Supervisor feels the man has been trained in school to perform each task effectively. Such response enables determinations as to degrees of under-training, adequacy of training, or over-training. The biographical data sheets provide additional in-put

for final analyses of the responses. They indicate the Supervisor's rate, the graduate's assigment, the length of time the graduate has been on the job, any intervening training he might have received, or any activities which might have delayed commencement of duty.

The Commanding Officer's critique offers a verification of the Supervisor's opinion of performance and adequacy.

The open-ended question for "missing elements of training" provides an opportunity for graduates and Supervisors to objectively enumerate training needs not being met and which appear to them to be essential for productive field performance.

RESULTS; INDICATIONS; IMPLICATIONS

The decision was made to initially survey each of the schools with follow-up surveys every six months. To date, all Class A Schools and subordinate tracks have been surveyed. The survey population consisted of 678 graduates, their Supervisors and Commanding Officers. Percentage of returns averaged 59%.

A first analysis indicates major AREAS of Over-training, Adequacy of Training and Under-Training. This gives an immediate indication of over-all Adequacy or Inadequacy of training.

Within these major AREAS, SPECIFIC task are then examined in light of training requirements.

Once over-all training effectiveness is determined, an in-depth study of both these categories, AREAS and SPECIFICS, reveals many aspects for consideration. The following are some ramifications of the surveys. It is interesting to note that even though the same evaluative vehicle was used, feed-back was different, and idiosyncratic of the individual school or track. A review of major AREAS shows generally that:

(1) if, as in the Electronics survey, the majority of responses indicates Adequate Training (most tasks are performed very often and trained well), then the over-all implication is that the newly instituted ET program is indeed accomplishing its mission. And this is particularly significant as an evaluation of training effectiveness since, as previously discussed, the Electronics Class "A" curriculum at Governors Island is totally a "hands-on" training experience. The change of policy to need to know, from nice to know, has succeeded, overcoming original antipathy. The ET program is teaching to CG-311, and concomitantly, CG-311 does define what an entry-level technician needs to be able to do in the field. It is interesting to note also, as supportive data to the field survey results, that students trained in the new curriculum faired better on the Coast Guard servicewide exam, which rank orders qualified candidates for advancement, than those previously trained under the older system in four out of six test areas (Administration, Safety, Electricity & Electronics, Solid State Theory), and less than 1 percentage point below previously trained students in the other two areas

(Test Equipment and Technical Maintenance), even though the old curriculum was heavily loaded with "theory" in contrast to the hands-on application in the new curriculum.

(2) if, as in the Gunnersmate and Damage Controlman surveys, results indicated some areas of gross over-training, investigation revealed that graduates' and Supervisors' rating of frequency of performance of task and adequacy of training ratings vary according to duty station. A Damage Controlman stationed on land does not perform tasks related to anchor windlasses and anchoring equipment, for example. Consequently, he rates these tasks as over-trained and suggests eliminating or drastically reducing training. However, the odds are, if he remains in the Coast Guard, he will be assigned to different billets where he must use these skills. Conversely, a Damage Controlman assigned to a floating unit would rate woodworking and building maintenance similarly.

(3) if, as in the Gunnersmate Survey results, a graduate is technically trained for a shipboard station and he is assigned to a shore station, his evaluation of training will be affected. If the facilities on which he was trained are not available for use, his responses will fall into the over-trained category. However, it must be borne in mind that at a future date he may be transferred to a duty station where the facilities are available, and at that time, he must have had some "overtraining" or his recall will be insufficient without having additional on-the-job training. Also criticality of tasks must be considered. In the event of a national emergency, that which is considered "over-training" in a peace-time Coast Guard might readily become "under-training".

Reference to SPECIFIC tasks indicated that, in ET training, the term "THEORY" which had been the "catch-all" of needs, disappeared, and specificity of training requirements pinpointed actual additional training requirements in the revised ET program and the three tracks. One important adjustment to Electronics Training which resulted from this definition of needs was the addition of maintenance of Communications equipment to the Loran Track curriculum. For an Electronics Technician at an isolated Loran Station, knowledgability of Communication equipment maintenance was vital. The practicality was evaluated and the curriculum expanded. Once correlated with the graduate's assigned duty station, the surveys indicate which particular training needs require strengthening as the thrust of his technical responsibilities at that station vary from the core program at the Training Center. They indicate a possibility of more accurately teaching to placement. They also supply information as to what a graduate needs when, even though an apprentice, he is assigned to a billet requiring a much more experienced man (for example, assigning an E-4 to a small ship wherein he alone is responsible for advanced as well as elementary duties).

SPECIFICS, as well as AREAS studied, also provide in-put for review of "in-house" Training Center testing. If categories are substantially weak in the field, perhaps "in-house" testing was an inadequate predictor. Perhaps the tests were not accurately measuring the man's

progress as he was being trained in particular skills. Perhaps a re-evaluation of testing methods is suggested.

A study of the "MISSING TRAINING ELEMENTS" question, which often expands itself to encompass aspects not strictly considered training needs, provides important feed-back for Training Center - Field communication and inter-play.

The request for more leadership training appeared thematically throughout one school survey. This is a problem which cannot be primarily addressed as a training need in our system. The sole mission of CG Training Center Governors Island is to produce an entry-level technician - and leadership is not indigenous in apprenticeship, although it is inherent in a military ethical standard. It is interesting, however, to try to discover a way to infuse this intangible moral element into vocational training without actually teaching it, as there is no provision for such instruction either in time or in the curriculum. Yet it is essential for advancement.

This area is now being investigated. We plan to use as an appendix to future field surveys a Military Performance Questionnaire. Both the graduate and his supervisor will be asked to evaluate how frequently tasks relative to military aspects of the Coast Guard environment are performed in the course of a graduate's initial duty tour and how adequately the emphasis was placed on this aspect during technical training. While this segment of militia is not taught per se at the Training Center, it must be learned through emulation. As instruction must be effective for productivity, so must be the Instructors, for their impact must not only be on a technically trained man, but on a <u>military</u> technically trained man.

The utilization of the initial job task analysis field survey provided an evaluation that training in our Class "A" Schools is adequate and effective. Areas of change which were indicated have been implemented. The appearance is that <u>at the present time</u> there has been achieved the delicate balance of productivity: the product is adequate for the degree of training, the amount of time expended and the training cost incurred. However, the politics of experience has showed there is rarely a constancy in a changing society and the Coast Guard is no exception. The advent of a lowered entry score for school admission, albeit exit standards remaining the same and what this implies, calls for another evaluative examination as does curriculum changes implemented after the initial field surveys. A process of periodic re-surveys, to measure training adequacy must be the barometer for productivity.

CAN TRAINING PRODUCTIVITY BE MEASURED?

Productivity measurement requires an analysis of input, process, and output. (See Figure 1). On the input side there are: (1) Man-power needs of the field; (2) Job-task inventories (JTI); (3) Established job performance standards; (4) School entry standards. The manpower needs translate into the number of students which will be enrolled in training during any one period. JTIs and Job Performance Standards dictate the training objectives which must be met.

The training process consists basically of methodology of training, length of training, and resources, manpower and funding.

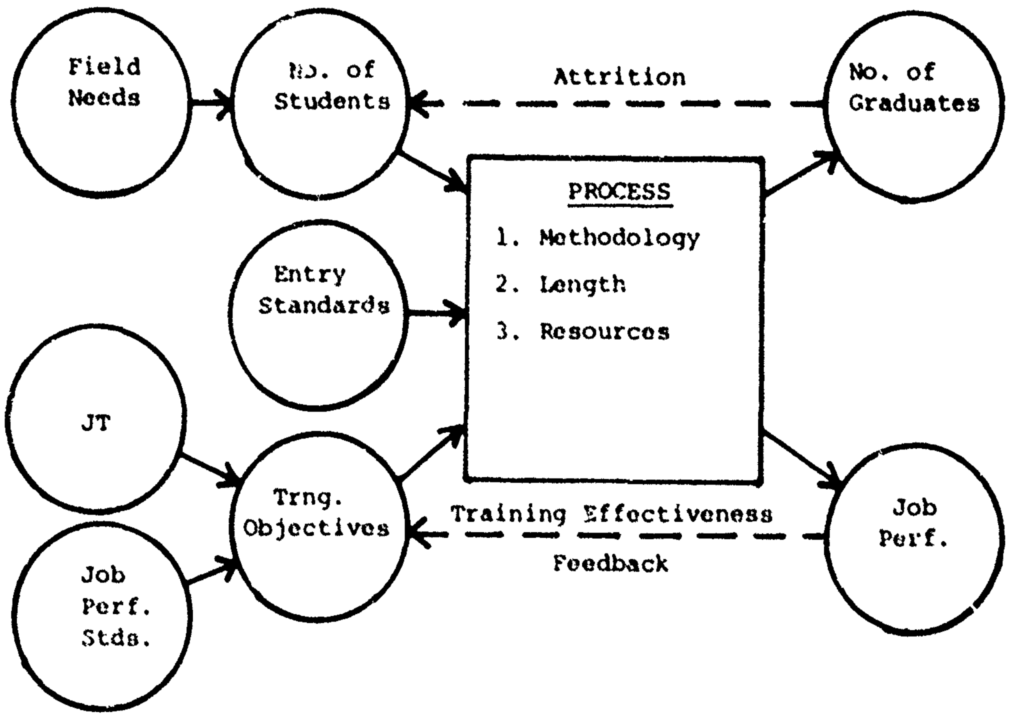


Figure 1

The measurable output consists of the quantity and the quality of graduates. These, too, have cyclical effect on input, where there is shortfall or overabundance of either.

If aspects of any three change, the productivity will change, either in quantitative or qualitative results. The ability to measure effectiveness is the key to measuring the productivity of the training system as we have seen. But can it be done? We believe so, although we have not yet found the right formula to fit our needs.

At Training Center Governors Island, we have seen that by converting our various curricula to performance-based instruction, we can reduce the

length of training and still maintain effectiveness. The resulting personnel cost decrease seems to have balanced the increase in capital costs required in a total performance based system. But what happens if changes external to our control occur? Any of the input variables may change which will have a change impact on the training process if the output is to remain constant. Requirements for change in the process will also impact on the output.

For example, last year the Coast Guard reduced it's class "A" School entrance standards for all but a few ratings. This move was intended to provide a greater opportunity for more people to be able to learn a technical skill and progress toward a Coast Guard career. Although school entrance standards have been lowered, job entrance standards remain unchanged. This means that some aspect of the training process has to be adjusted in order to maintain the necessary quantity and quality of graduates. Frankly, we don't know what the overall impact is, although we have accepted the fact that the length of training will probably increase and that the attrition rate may rise.

There is also evidence that the Coast Guard may undergo a drastic decrease in support billets over the next few years. That indicates a cut in training resources, even though the input and output needs will remain unchanged. There will certainly be an impact on our training productivity even though other in-house adjustments to either methodology or length of training are made.

Some of our ratings are currently undergoing job task inventory. We welcome that. But as the job tasks may change so must our training processes. Zero-based budgeting requires the necessity to measure productivity of every program conducted. If we were to be asked what we could produce if our FY79 budget was cut 80%, what could we answer? That we would produce 80% fewer qualified graduates, or the same number, but each only 80% qualified? We need a more definitive method to provide some definitive answers. We not only need to quantify our current productivity, but we need to be able to see how a variety of changes will impact on that.

PRODUCTIVITY BENCHMARKS

The first thing we have to do, after determining the effectiveness of our current training, is to establish some realistic benchmarks about which we can make realistic judgements. The first benchmark necessary is the job-performance standard. In our system, job performance standards are established by the program manager or subject-matter expert. The standards have to be translated into quantitative values so that comparisons can be made with in-training performance test standards. If valid, performance test results should be relatively accurate predictors of on the job performance. In our field survey system, a verdict of adequacy (neither over nor under-training) would confirm that our performance test standards are equivalent to job performance standards. For example, in the Electronics Fundamentals Section of the ET curriculum, field surveys showed the training to be adequate. 3.900

performance test scores were than analyzed to determine the average scores in each module. These average scores have then been established as the quantitative job performance benchmark for those tasks trained. To make this benchmark meaningful in productivity measurement a relative cost to achieve it has to then be determined.

Juran's "Quality of Conformance Model" (See Figure 2) may help to establish and then confirm the job performance benchmark in terms of costs. According to Juran, (1) the more one strives for perfection in a product, the higher will be the cost of producing one good unit of that product relative to a conformance standard. A state of diminishing returns will be eventually approached. On the other hand there will also be costs incurred for an imperfect product (one less than 100% perfect), which Juran calls "cost of failure". The cost of failure increases as the degree of perfection of the product decreases from 100% perfect. In essence, high costs are incurred at either extreme of conformance to a standard. Using these principles, a training manager should be able to plot the costs of successful training against the cost of failure in job-performance to determine total costs relative to productivity. Conformance or quality standards based on job performance standards could then be compared with total quality costs and adjusted as necessary to achieve optimization.

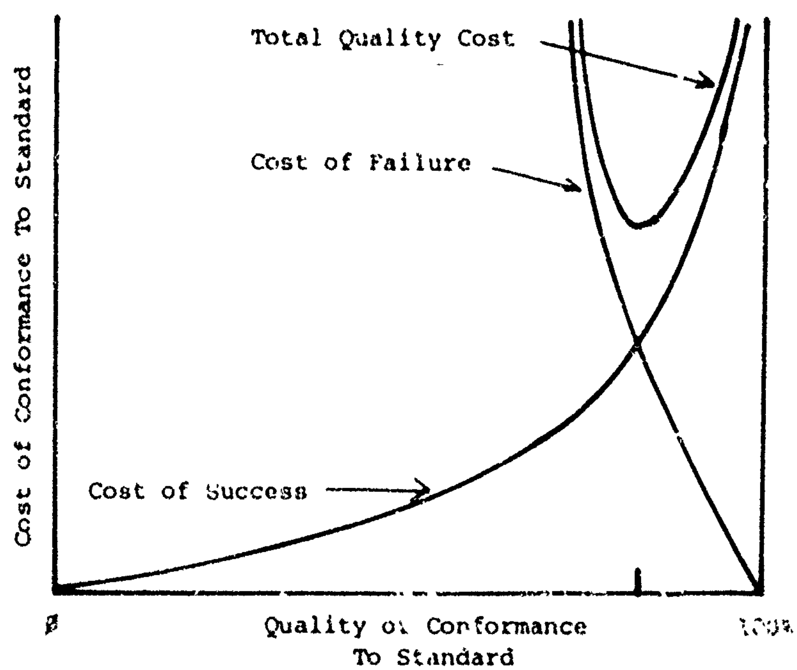


Figure 2

(1) QUALITY CONTROL HANDBOOK by J. M. JURAN.

For example, assume that a conformance standard of 70 on a scale from 0 to 100 is considered as qualifying in any given situation. A training cost to meet that standard can be computed based on existing data. An adequate feedback system could then show the relative cost of failure, that is, the training deficiency in the range of 71 to 100 on the same scale. If it is found that the relative cost of failure is excessively high, then apparently the standard is too low and should be adjusted upward. If the relative cost of training is too high, then the standard should be adjusted downward. The relative cost of failure will be an extremely difficult value to define and will probably have to be expressed only in terms of a percentage of under-training determined from effectiveness surverys.

The standard would then fall into one of three zones described by Juran on a total quality cost curve: a "zone of improvement" (standard is too low), a "zone of perfection" (standard is too high), and an "optimum zone" (standard is just right). (See Figure 3).

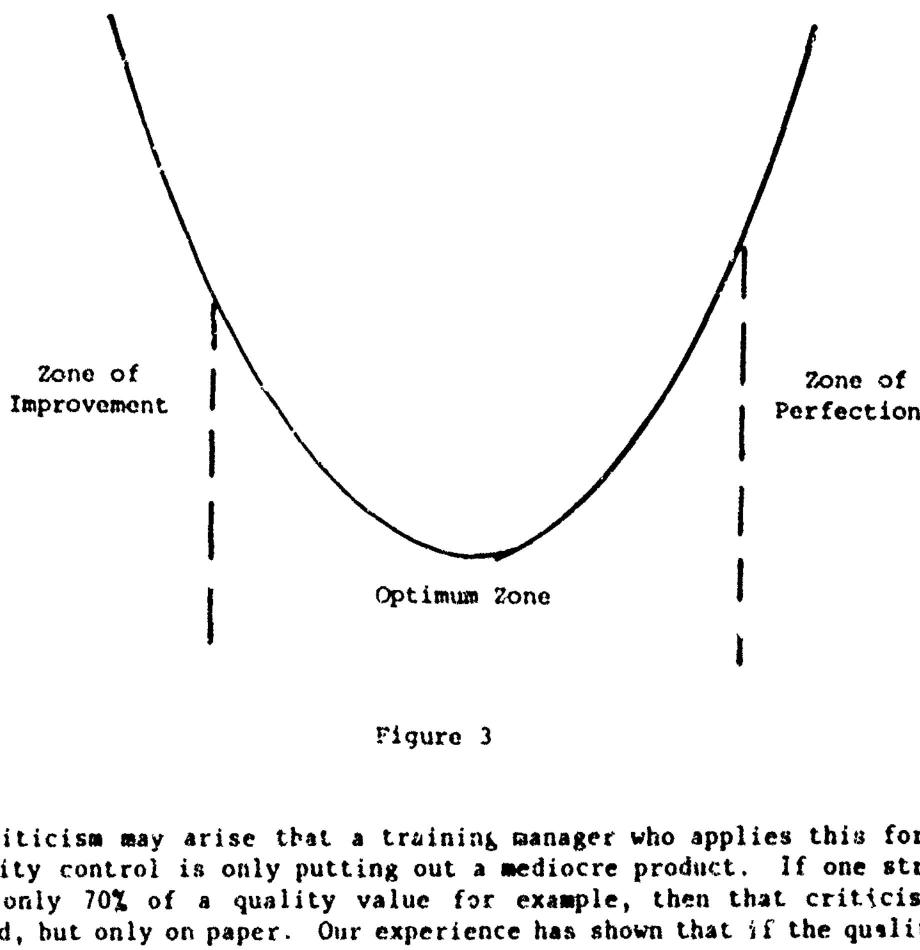


Figure 3

A criticism may arise that a training manager who applies this form of quality control is only putting out a mediocre product. If one strives for only 70% of a quality value for example, then that criticism is valid, but only on paper. Our experience has shown that if the quality

standard is realistically based on the best data available, a great portion of students will achieve beyond that standard provided the motivational environment is right and no other restraints are placed upon them. To paraphrase an old duplicate bridge axiom: "Striving for the best possible result will result in failure most of the time. Striving for the best result possible will result in success most of the time." Granted, there are those areas where perfection in training is an absolute must for job entry. The astronaut training program is an example. Perfection here is necessary because the cost of failure is so great. The program manager has to accept this cost if the program is to continue. In our experience though, that condition is the exception rather than the rule.

The Juran conformance to standard model can also be used to establish and confirm a realistic pass-fail threshold. For instance, the value which is represented by the line of demarcation between the optimum zone and the zone of improvement on the total quality cost curve might be considered to be a realistic pass-fail threshold once the quality standard has been determined.

The next benchmark for measuring training productivity is a realistic attrition rate. This certainly relates to an established conformance standard. By applying Juran's principle here, as well, an optimum zone can be determined. Since our task as training managers is to provide a given quantity of qualified graduates to meet field needs, the <u>ideal</u> attrition rate is zero. But from a productivity viewpoint that is not realistic. If the conformance standard is set so low that every student passes, there is a risk of higher costs of failure in job performance. Conversely, if the standard is set too high, the quality needs will be met but the quantity probably won't be. Therefore, the pass-fail threshold may have to be adjusted accordingly, even though slightly higher than optimum cost of failure may result. This circumstance may be offset by additional follow-on OJT to reduce cost of failure somewhat.

Cost per student trained is the third benchmark. Student time, instructor time, training equipment, and operating costs are elements which make up this benchmark. It is the probably the most critical of the three because productivity in any management area has to be related to cost as can be seen in Juran's model. It interrelates directly with the other two benchmarks because they are expressed in terms of quality and quantity costs. These three benchmarks, job performance standards, a realistic attrition rate, and cost per student trained are the basic keys to training productivity measurement. We call them benchmarks because they are just that: starting points on which we can base a training productivity measurement. The benchmarks must have some flexibility and should be adjusted as necessary to eventually "bracket in" on a firm standard. Each, of course, must be confirmed by results of training effectiveness surveys. Once established, these benchmarks should provide starting points to enable us to solve some of the problems which were discussed above. Please keep in mind that we have not yet come up with a workable formula, but we are continuing our research. We will explain generally how such a productivity measurement system may be applied.

APPLICATION

Consider some of the problems mentioned previously. The first of these is the effect of reduced school-entry standards with no offsetting reduction in job-entry standards. With reference to the input-process-output model, (Figure 1) the only adjustment which can be made is in the process. But first we have to see what our current productivity is before we can measure the effect of change. Let's look at our ET curriculum as the best example.

Based on field surveys conducted prior to this change, we know that our current ET training process is meeting both of the field quantity and quality needs. We also know the number of student hours, instructor hours, cost of training equipment, and operating costs involved to meet those needs. From analysis, we have established a meaningful performance standard and pass-fail threshold. Our attrition rate has leveled and has remained fairly constant at 5%. We also know the training cost per student. These are our initial benchmarks.

Let's then consider what the results of lower entry standards might be. We suspect that it is a greater number of students with some reduced learning abilities, probably lower reading and mathematic skills. Next, consider what the impact on productivity will be if no changes in process are introduced. Without reducing job performance standards (benchmark 1), it is obvious that a higher attrition rate (benchmark 2) will result. Consequently, we will fail to meet the number of graduates required by the field. If performance test standards are reduced in order to maintain the lower attrition rate, the risk of on-the-job failure will result. Therefore, a change in process is mandated. We could change the methodology, increase the length of training, or increase resources. Each will result in some cost increase (benchmark 3), thereby effecting productivity. We believe that the current methodology, (performance-based instruction) is ideal for this situation, although we are considering adding some self-study enrichment programs which may assist. An increase in resources, except for student time, is probably not the answer. Therefore, the logical adjustment is to increase the length of training. In essence, the slower learning student will be given more time to learn. This will result in a proportional increase in cost per student trained.

What does this mean in terms of productivity? A greater course length will result in a relatively higher training process cost. Is that productive? If it meets the objective of the school-entry standard reduction (to give a greater opportunity for more people to learn a skill) and still satisfies field needs, the answer is yes. The next step is to conduct a field survey to determine the job performance of those graduates who had entered training under the lower entry standards. This will serve to measure the effects of the reduced school-entry standards on job performance, if any, to confirm that the adjusted training was effective, and to confirm the adjusted cost of training benchmark.

Another problem which we expect to face soon is a reduction in instructor billets. We don't know how many billets we might lose but we hope to be able to make some realistic projections of it's impact. If our envisioned productivity benchmarks are valid and confirmed we should be able to pinpoint those courses in which we can effect the resource reduction and what impact that will have. It may require either change in methodology or in course length or in perhaps some combination of both. If it appears that these changes may result in a reduction of training quality or quantity of graduates, then that too should be evident. In any event, the impact of a resulting cost of training decrease which should result from reduction of resources can be revealed. Is the sacrifice of some degree of quantity or quality of graduates in favor of lower cost of training productive? Perhaps it is, if the relative cost of failure on the job is not too great.

What about zero-based budgeting? A productivity measurement system is an absolute must. When a manager can show in quantitative terms the effects of resource changes on productivity, his budget estimates have a firm foundation. Many of us at our levels have not had to zero-base our budgets yet. But it will eventually come down to us. We will have to analyze every element of our training programs to justify how we are conducting them. Well defined standards, a realistic attrition rate, and specific cost data related to these will aid us in that justification.

SUMMARY

In this paper we have presented issues which are undoubtly not new in management. Nor is our approach to resolving them necessarily unique. Perhaps it is innovative only in our eyes - the eyes of short-term training managers who will soon move onto some other field of endeavor. Perhaps naively, we feel that it is the most practical approach to fit our needs at Coast Guard Training Center Governors Island.

The problem presented is manifold. Is our training effective? Is our training system productive? What are the impacts of a variety of changes which may occur? How can we adjust to those changes and still be able to meet the quanity and quality of graduates needed by the field? What are our benchmarks?

We have so far found only a partial solution to this complex problem. More work is yet to follow. We have found that we can measure our training effectiveness. This measurement allows us to validate performance testing and to insure that these tests adequately predict on the job performance. Through analysis of the feedback data and performance test results, a job-performance bench mark can be established.

Further, through the same analysis, a realistic pass-fail threshold is established. This value allows us to "bracket in" on a reasonable attrition rate - one that insures that the quanity of graduates is responsive to field needs and at the same time minimizes the cost of failure on the job.

Finally, we can develop a cost per student trained benchmark which encompasses all costs involved in the training process. This is the pivotal benchmark because it allows us to relate the quality of training and quantity of graduates to a cost value. Each of these benchmarks must be considered in the productivity measurement system which we envision. The next step will be to develop a formula which can be applied to answer any of the questions posed above. We think we are on the right track, but it is too early to tell.

The main rationale behind this paper has been not only to explain what we have done and what we hope to do, but to appeal to the military training community for assistance. We would greatly appreciate any feedback we can get.

FEMALE MILITARY PERSONNEL UTILIZATION AND COST EFFECTIVENESS

MARILYN E. HARRIS AND LINDA PAPPAS

ABSTRACT: *This paper presents a comprehensive research approach toward improving the full utilization of personnel resources in a cost effective climate, specifically improving the attitudes toward increasing the utilization of women personnel (enlisted and officers) in maintaining military effectiveness. The authors in reviewing the relevant literature of both private and military sectors, acknowledge the "blocks" to utilization of personnel, as well as limited evaluation research technology and experimentation. Three categories of "blocks" as discussed in relation to the military, they are: Legal, tradition and values, and sex differences. The applied training model (treatment intervention) focuses on attitude change based on content learning and performance analysis where effectiveness is evaluated pre, interim and post periods. The applied training establishes leadership development for enlisted officers, males and females in single sex groups and mixed groups, on 1440 subjects. The success is measured in cost savings to the military and compared to a 5-20% savings for the private sector.*

FEMALE MILITARY PERSONNEL UTILIZATION AND COST EFFECTIVENESS

Traditionally, the military establishment of the United States, has been a masculine domain. For the greatest part of this nation's history, women have not been permitted to serve in a military uniform except as nurses. Indeed, serious interest in defining women's role in the armed forces did not awaken until World War II, and it was not until 1948 that women achieved permanent military status. Thus, the formal association of women with the armed services is a relatively recent phenomenon. The changing role of women in the military establishment of the United States in large part has mirrored their changing role in American society; it has recently been influenced by military necessity.

HISTORY AND DEVELOPMENT: A Brief Overview of Women in the Military[1]

Prior to World II, the United States Armed forces were almost exclusively male. While legendary women warriors such as DEBORA SAMPSON ("Robert Shurtleff") in the Revoluntionary War, LUCY BREWER ("George Baker") in the War of 1812 and MOLLY PITCHER did exist these were indeed exceptions to the rule. Other women, as civilians, assisted the military in such capacities as nurses, cooks, laundresses and other acceptable feminine pursuits.

[1] The major portion of the background section is from the Brookings Institution study Women and the Military, Chapter 2.

In the wake of nineteenth century industrialism, American Women developed skills that were to become increasingly relevant to the military. In fact, women dominated some occupations (for example, secretaries and telephone operators) which, with changes in military technology and organization, had come into greater demand by the armed services by World War I.

During World War I some 13,000 women were enlisted in the Navy Department, However, after World War I they were demobilized. Thus, the few remaining women in the armed forces were found in the nurses corps.

Between the two World Wars little interest in women in the military existed. Although two plans [1] were designed that addressed women in the military, neither plan received sufficient support.

World War II can justifiably be viewed as the turning point in the history of women s quest for military status. Large numbers were involved -- a total of about 350,000 women served in the four military services. And although the vast majority were employed in health care, administration, and communications, women demonstrated their competence in virtually every occupation outside of direct combat, including airplane mechanic, parachute rigger, gunnery instructor, air traffic controller, and naval air navigator. It is also worth noting that some 800 women served as Women's Air Forces Service Pilots (WASPS). Although never accorded full military status, they ferried all types of military airplanes, including combat aircraft.

Women's role in the Second World War was far more significant than is suggested by the brief overview presented here. Perhaps the ultimate compliment paid to the American women who served was offered by Albert Speer, Adolph Hitler's weapons production chief, to now-retired Lieutenant General Ira C. Baker, an Army Air Force commander in Europe during World War II:

> How wise you were to bring your women into your military and your labor force. Had we done that initially, as you did, it could well have affected the whole course of the war. We would have found out, as you did, that women were equally effective, and for some skills, superior to males.

[1] One plan was developed under the auspices of Anita Phipps, the Army's Director of Women's Relations; the other plan was directed by Major E.S. Hughes.

With the conclusion of World War II came a large and rapid demobilization. Further, the authorization for the WASPS was set to lapse in 1948. Therefore the women who decided to remain on active duty after the war were in a precarious position.

In 1948 the Women's Armed Forces Integration Act was passed. Hence, women's role was clarified; they were given the opportunity to pursue a permanent military career. Many factors precipitated the enactment of this legislation; one major concern was that the armed services, soon to be without benefit of conscription, would have difficulty in meeting their recruitment needs by voluntary means. Though signifying a major breakthrough for women, the 1948 legislation also sowed the seeds of sex discrimination that was to persist for two decades. Specifically, the act imposed some major limitations on women in the areas of recruitment, career opportunity and dependency status.

With the decision in 1970 to end the draft, the United States embarked on a venture unprecedented in any nation's history: to field a military force over two million strong relying solely on volunteers. Could enough men be found, willing and able to volunteer, without exorbitant additional costs, and without compromising the quality of military manpower?

The combined impact of the end of the draft, the Equal Rights Ammendment debate, the feminist litigation and the changing social attitudes affected the women in the military in five major areas:

- The number of women in the military has increased substantially;
- Personnel policies have been changed (e.g. aviation training, command billets, wider job range etc.);
- The number of specialties open to women has increased dramatically;
- The proportion of women assigned to nontraditional jobs has increased; and
- The influx of women into the military has influenced the socio-economic composition of the enlisted ranks.

Inspite of the increased numbers of women in the military and their expanded roles, there still exist factors that inhibit the maximum utilization of women in the military.

INHIBITING FACTORS

The factors that inhibit the maximum utilization of women in the military may be grouped into three major categories:

Legal In 1969 the Secretary of Defense and the Chiefs of Staff of all the Services signed the Department's Human Goals statement. This statement declared that the defense Department would strive "to make military and civilian service in the Department of Defense a model of equal opportunity for all regardless of race, sex, creed or national origin ...[Commanders Digest 15(8), 1974]. However, some laws exist that prevent the full implementation of this policy. Specifically legal restrictions are imposed on the military by Sections 6015 and 8549, Title 10 of the U.S. Code. Section 6015 prohibits the use of women on Navy vessels, other than hospital or transport ships. Sections 8549 prohibits the assignment of women to aircraft engaged in combat missions.

[1] CONTRARY TO WIDELY HELD BELIEFS, the major restrictions on the recruitment and functions assigned to women in the United States military establishment are not explicity incorporated in federal law. To be sure, few opportunities in either the Army or Air Force would be closed to women if the statutory provisions governing the utilization of military women were literally interpreted. More limiting are the set of policies established by the military services themselves based on their own interpretations of the national will as expressed through Congress. Together, these laws and policies relegate women to minor role.

Tradition and Values The number of female applicants to the military demonstrates their willingness to serve in what traditionally has been a male career arena. However, reluctance appears to exist to select non-traditional jobs within services. A private sector study (Integration of Females Into Male-Oriented Jobs: Experience of certain Public Utility Companies, University of South Florida, 1976) suggests similar findings in the private sector with respect to blue collar jobs. It was hypothesized that women frequently had the opportunity to acquire the knowledge and skills required for non-traditional white collar jobs, but did not have the opportunity to acquire requisite skills for blue collar jobs. Another factor appeared in experiences with blue collar jobs. Woman did not have as much intrinsic interest in non-traditional blue collar jobs as in the non-traditional white collar jobs.

[1] Martin Binkin and Shirley J. Bach - WOMAN AND THE MILITARY The Brookings Institution, 1977 - p.30

The factors of tradition and values, though very difficult to measure, play a significant role in obtaining (or hindering) maximum resource utilization of all military personnel. Women's role in the Armed Forces will ultimately depend on the extent to which National Institutions ---- social, political, judicial and military ---- are willing to break with their past ---- a past reflecting a persistant pattern of male dominance.

Sex Differences An often quoted complaint by women's groups is that even the most sophisticated and sympathetic male managers cannot easily shed deeply ingrained attitudes regarding the proper roles of women in society. These attitudes creep into many decisions in which a person's sex is not an obvious factor. Very often these are not straightforward decisions involving favored treatment of a man over a woman, but rather decisions which result in a specific treatment of women, without any regard to how a man would be treated in identical circumstances.

Various traits of a "perfect man" and a "perfect woman"(have been) established through sex-role stereotypes. Research has indicated that there is a positive relationship between the profile of mental health for an adult male and the general profile for a healthy adult, sex unspecified. "Healthy adult female behaviors, then are seen as less socially desirable and less mentally healthy than the behaviors of healthy adult males".[1]

The general increase of women in the labor force and their inroads into traditional male employments has led to a gradual change in society's attitude concerning women's capabilities. Cases of women performing satisfactorily in traditional male employments are no longer newsworthy. Yet these changes have filtered slowly through to the armed services. The acceptance of women in the military has been overshadowed by the controversy surrounding the possible acceptance of women in combat.[2]

The issue of combat is indeed a complex one and a complete discussion of combat is beyond the scope of this paper. However, with respect to tradition, George Quester has noted in his article "Women in Combat" that "A nation forced to send its women into combat must be the underdog, the nation that has been threatened, the nation that cares the very most about the justice of its cause."

[1] Ibid

[2] Women (and Men) in the U.S. Army: A study in Optional Utilization. Michael John Castle, Naval Postgraduate School, Monterey Calif., December, 1976.

One vivid example of how deeply stereotypes pervade our thinking is described in a study conducted by Inge Broverman. The study involved various kinds of psychotherapists, psychiatrists, psychologists and social workers. The Broverman investigators developed a list of 122 pairs of traits (e.g., "very active vs. very passive"), and gave these lists to the therapists. The therapists then employed a semantic differential format and broke out each trait into a seven point scale (i.e., extremely active, moderately active, slightly active, neutral, slightly passive, etc.). Having completed this breakout, the therapists then were to indicate on the 122 scales where they thought a healthy adult would fall.

Next the therapists were to go through the scales again and indicate where a healthy male and a healthy female would fall.

It is important to note that the therapists were not asked to describe men and women as they are but as the experts thought these men and women ought to be. "The results (with no significant difference from male and female therapists) were quite simple. A healthy adult is a healthy male."[1] The stereotype is very pervasive in societal thinking, especially when considering the qualities and capabilities possessed by a person of a specific sex to perform a non-traditional job.

It is important to examine further how sex differences may affect individual capabilities, group performance and image in relation to military effectiveness. Complex factors such as discipline, leadership, training, societal influence and group relationships all bear on efficiency.

It is clear that personnel quality (measured by educational level and general intelligence and apitude) of women in the armed forces improves the overall level. Individual differences far outweigh sex differences. That is, the differences within males and within females on ability variables are far greater than the differences between the means of the sexes. A second finding is that the literature contains many contradictions relative to sex differences. Results are often confounded by unconscious bias, the actual task used, and overgeneralization from one sample to the population. Although findings tend to be inconsistent (see, for example, Bond & Binacke, 1961; Maier, 1970; Megargee, 1969), it has been reported that males and females are similar in leadership ability (Day & Stogdill, 1972), problem solving (Mathews, 1972), cooperation and competition (Lirtzman & Wahba, 1972), and potential capability (Bass et al., 1971). However, women are not always accepted. This may be due in part to the over-

[1] A Male Guide to Women's Liberation, Gene Marine, Avon Books, New York 1972

emphasis of group differences, and not to evaluation based on individual qualifications. The resulting resistance often has been manifested as lower pay (U.S. Department of Labor, 1971), lower positions in the organization (Fidell, 1971), and a general under-utilization of women in the work force (Kootz, 1970).

Research suggests that as long as women are in the minority, men fulfil their own need to project the male image. This would tend to isolate women, keep the male group in conflict with them, and thus reduce over all group productivity. Integration studies at Yale and Princeton Universities found generally that, while the ideal mix was not surprisingly half and half, social problems were less likely to develop when the ratio of men to women was lower than three to one. Above that threshold, according to the researchers, some women tended to assume a "superwoman" role and to make more male friends than they normally would, while the men tended to socially reject them as inferior.[1] Interpretation of the determinants of behavior in combat suggests that the introduction of women into fighting organizations or seagoing units would have less disruptive effects on solidarity.

The image abroad would probably attract little attention anywhere, unless it resulted in a dramatic shift in total sex composition or unless it was accompanied by an unprecedented integration of women into U.S. fighting units. As the discussion above indicates, a healthy measure of uncertainty remains about greater female participation would affect the factors in sex differences noted.

COST EFFECTIVENESS: Through increasing Women in the Military

While social forces must be considered in the development of national policies with respect to women in the military, so too must a dollars-and-cents, cost-effectiveness. If an increase in the proportion of women in the military could lead to lower costs without sacrificing effectiveness, the United States is paying more than is necessary to field its present military forces. The financial implications associated with a change in the mix of men and women in the armed forces shows not only that the cost differential previously associated with a higher expected personnel turnover rate has been largely eliminated, but also that if one-time costs for separate facilities are incurred, they would be offset, at least in

[1] James H. Thomas and Dirk C. Prather, "Integration of Females into a Previously All-Male Institution, "Proceedings of the Fifth Symposium on Psychology in Air Force (United States Air Force Academy, Department of Behavioral Sciences and Leadership, April 1976), pp. 100-01.

the short run, but annual savings that would result from supporting a smaller dependent population. Furthermore, the larger costs that might accrue because women would be less productive on the job than men for reasons of pregnancy or illness are now more than offset by the greater tendency of men to have disciplinary, drug, and alcoholic problems.

All in all, the force of admittedly scanty evidence is the changes in the sex composition of military services would lead to changes in total costs. Over the long term, however, such differences in costs would probably narrow as more women enter the military services - and one that is likely to attract more attention - is the prospect of being able to maintain desired quality standards among volunteers.

Generally then the factors in cost effectiveness related to quality military performance are both social and military and require further examination and careful documentation of all costs over a specific period of time.

The importance and value of full recognition and equality for women in the military is being tested at all levels, in both traditional and non-traditional roles. For the most part, research has proceeded by focusing on the conceptual analysis of a single variable. Findings from both the military and private sector research on career success have established that manipulating one variable or one procedure does not provide career success in and of itself. In fact, what is typically required is a program that incorporates many components and procedures of career success in a systems approach.

Recent studies [1] identifying differences between female and male executives point to the importance of recognizing (1) the masculine perspective of supervisory and management roles and (2) the nature of both formal and informal relationships among corporate executives. This research is applicable to the military organizations which were built by men and for men, and are now controlled by men. The forms, rules and styles of behavior and communication among corporate supervisors, managers and executives grew from a distinctly male culture. An understanding of these variables is essential to career success. Recognition of the male culture also defines male involvment in any treatment designed to improve the work climate.

[1] The Managerial Woman, Margaret Hennig and Anne Jardim, Anchor Press/ Doubleday, N.Y. 1977

Successful efforts to improve female performance in the private sector have demonstrated the following facts. As women assume new roles, they must receive appropriate training in self management skill development and attitude change to prepare them for successful performance. Additionally, of great importance is the fact that women must learn the informal rules of organizational behavior and communication that are typically inherent in the male socialization process. Also related to the training and development of women is a parallel development of understanding and acceptance of change among male supervisors, peers, and subordinates.

It is acknowledged that there are complex problems existing relative to the full utilization of women in the armed services. Some of these problems will take further effort to resolve. However, we believe that even within the existing constraints greatly improved utilization of women in the armed services is possible.

THE PROBLEM: Fuller Utilization of Women vs. Military Effectiveness

The multifaceted problem is difficult to continue exploring "piecemeal" given the overall implications for military effectiveness. The armed services appear reluctant to address the issue at this time.[1] Therefore a comprehensive experimental approach is proposed in which the military system may be studied and evaluated given applied training treatment groups and appropriate control groups, which will identify contributions of each aspect of the system to the results at at least three intervals: pretraining, post training and and pre-combat, and post combat.

In summary the problem is extremely complex, involving a cross-cut of social and military factors. Two powerful social forces are in collision: the push for women's equal rights is in conflict with deeply rooted traditions that question the propriety of women under arms. That the body politic supports equal opportunity in principle is indisputable; however, the extent to which people will accept equality in pracitce, including committing women to combat, is less clear. Further the budgetary advantages of recruiting more women are at variance with the perceived risks to the U. S. national interest. The problem then is implementing principle without decreasing military effectiveness, and increasing the utilization of women personnel resources and increasing overall military effectiveness and maintaining cost effectiveness.

[1] Ibid, p. 110

THE RESEARCH IS FOCUSED IN THIS QUESTION:

Does applied training (as defined) affect the optimum integration of women into the armed forces and show cost benefits for a military effective climate?

THE PROPOSED RESEARCH APPROACH: OVERVIEW

The approach is a basic experimental-control panel survey design. It is tailored to studying the effect of applied training on military effectiveness when there is fuller utilization of women in the military organization. The design of the study is demonstrated in Figure 1. It is intended to cover a period of not less than 18 months. The projected study would consider the effect of the specialized training on twenty four groups (twelve enlisted and twelve officer groups) of 30 military personnel each drawn from integrated select units, where at least half of the units (treatment and control) would be going on to combat experience.[1] The research would study the effects of the specialized training in the all male group all female group and the two mixed groups of both male and female personnel as well as the effects on the unit from which personnel were selected at each site (all group members would be selected from one unit). Criteria for selection, classification and assignment of personnel, both men and women, would be important as would be the measures of effectiveness to be used. At each experimental site an equivalent control group would be identified and tested. The testing would consist of:

(1) An application selection test;

(2) Pretesting for base line data;

(3) Interim testing during applied training at the end of each phase (3 times)

(4) Post testing for growth as a result of applied training; and

(5) Follow-up testing at six months and twelve months.

[1] Officer group size maybe adapted in order not to cause impossible organizational configuration at the time. See p. 73 Women and the Military.

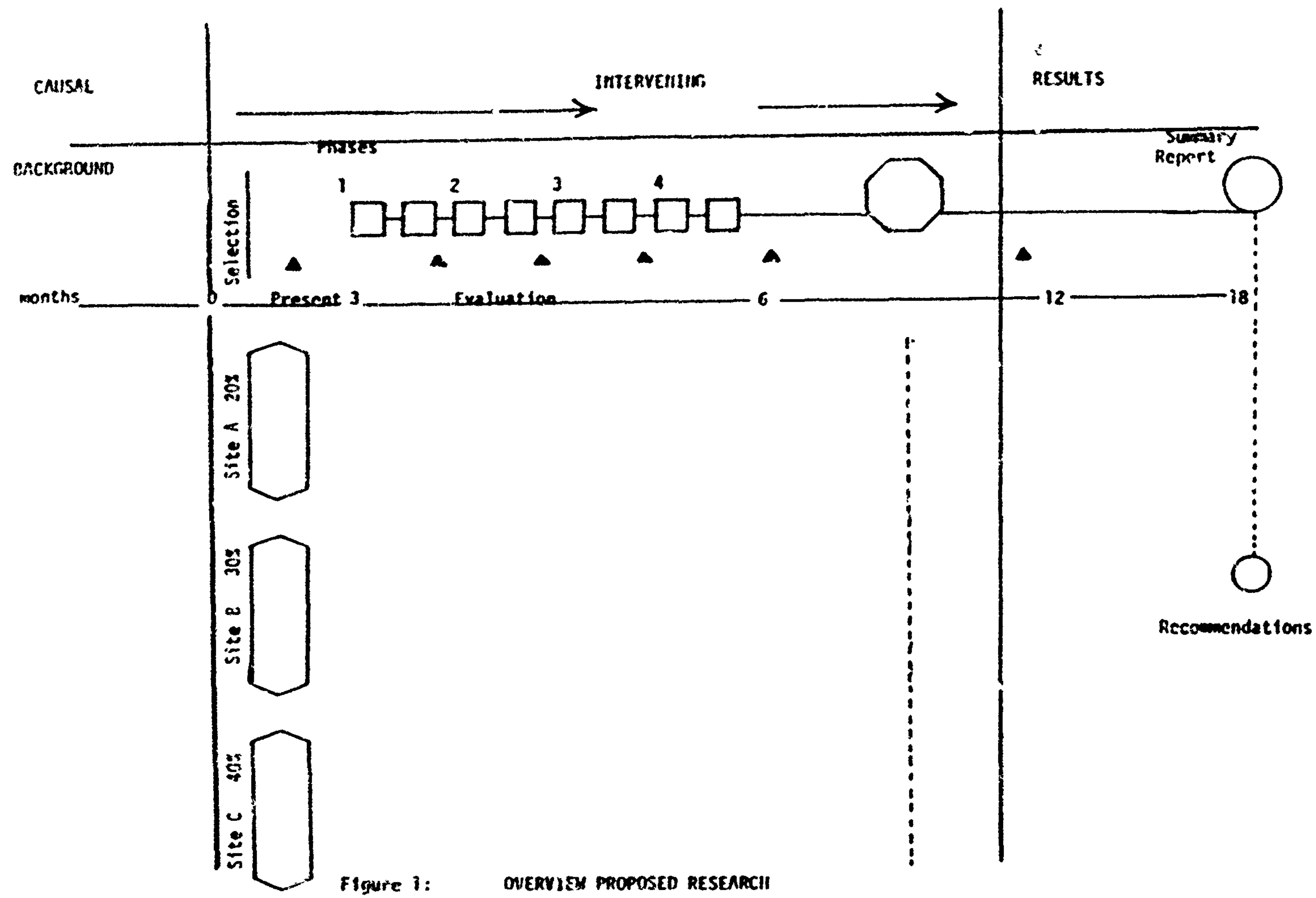


Figure 1: OVERVIEW PROPOSED RESEARCH

The follow-up testing would be to assess the internalization of new behaviors and attitudes in military effectiveness including results in combat experience, special training on performance and attitudes. It is recommended that this design be implemented at three sites where there could be experimentally arranged units of 20%, 30% and 40% integration of women into the military organization. The design would use 1440 subjects, 720 of whom would be control subjects. Figure 2 shows distribution at one site. This would be duplicated at the other sites for 30% and 40% integration.

TREATMENT CONTROL

	Type	Group	N	Group	N
	F	1_T	30	1_C	30
	M*	2_T	30	2_C	30
ENLISTED	M*	3_T	30	3_C	30
	M	4_T	30	4_C	30
	F	5_T	30	5_C	30
	M*	6_T	30	6_C	30
OFFICERS	M*	7_T	30	7_C	30
	M	8_T	30	8_C	30

* Mixed (15M-15F)

F = Female
M = Male
T = Treatment
C = Control

Figure 2: SAMPLE: Applied Training/Control Groups at Site 1 (20% integration of women.)[1]

[1] Site 2 (30% of women) and site 3 (40% of women) integrated into units.

TREATMENT INTERVENTION: Applied Training, Optimization and Adaptive Instruction.

A brief description of the applied training approach is in order to separate it from general teaching and training models in use.

The applied training approach treats the teaching-learning process phenomenon as a whole, recognizing the total environment and specifically the interrelationships between the content (competence), cognition and individual differences. It is assumed that what an individual subject does and can learn depends inextricably on what is already known. Thus in the learning setting, adaptive instruction, defines the underlying competence of the learner and builds from that point realizing what he/she needs to know in the content domain; identifying and developing the instructional structures; the stimulus experiences (exercises, simulations, role plays); and the steps for acquiring the specific behavioral competancies (skills). The applied training approach is an active, intense and highly interactive teaching-learning process providing the opportunities throughout for integrating the cognitive, behavioral and affective domains in each learning setting,consistently reducing the number of internal blocks to learning and change and increasing productive performance potential that is cost effective. The process learning is EAGT; where Experience is shared structured experience in the stimulus setting, Articulation of the experience, verbalizing behavior, affect and cognitions, Generalizing major learnings from the experience and Transferring through Identification, potential use in a novel situation. The model moves the learner in his/her own reality, moving from what they know, identifying the new structures for learning and then returning to a new reality, where the learner can adopt the new learnings. The applied training is designed in four segments to be held over a three month span providing opportunity for transfer to the real world and accountability for learnings upon return to the next training session.

The design of the applied training capitalizes on closing the "practice to use" gap, implementing the new learnings as soon after the experiences of "try out". To get a fair try out, participants plan before they leave the session how they will implement the new learnings in their day to day activity and when they return to the next session, the session begins by reviewing implementation and dealing with the problems before we begin new subject matter. Actually providing opportunities for try out of new behaviors is the beginning of the relevant attitude change.

The scope of this applied training design addresses these objectives:

A. Primary Objective

- Improving the attitude toward utilization of women personnel (enlisted and officers) in a military effective climate and specifically improving and increasing the utilization of of female personnel demonstrating cost effectiveness

B. Secondary Objectives

- Training women in appropriate military organizational behaviors that promote successful career patterns.

- Developing the self management know-how and leadership skills of female military personnel at pre-supervisory, supervisory, and management levels.

- Improving military personnel's sttitude toward successful job performance of women in new roles and non-traditional career ladders, as well as in combat.

- Improving both officer's and enlisted personnel's understanding of male/female socialization issues in relation to military job roles that impact on the military organizational structure and military effectiveness.

For the purpose of structural instruction and evaluation, performances objectives are specified for each secondary objective and are specifically addressed in one or more of the four modules of the applied training. The treatment intervention, the applied training program/process is the global experimental variable, independent in nature, and specified in the context of the training and evaluation. The dependent variables are identified in each secondary objective (skills, pattern, policies and practices; attitudes; and effectiveness) and are measures of success as a result of the training. Figures 4 and 5 describe variables and end results in relation to military effectiveness focusing the major intervening variables.[1]

[1] See Figures 4 and 5 for description of Variables

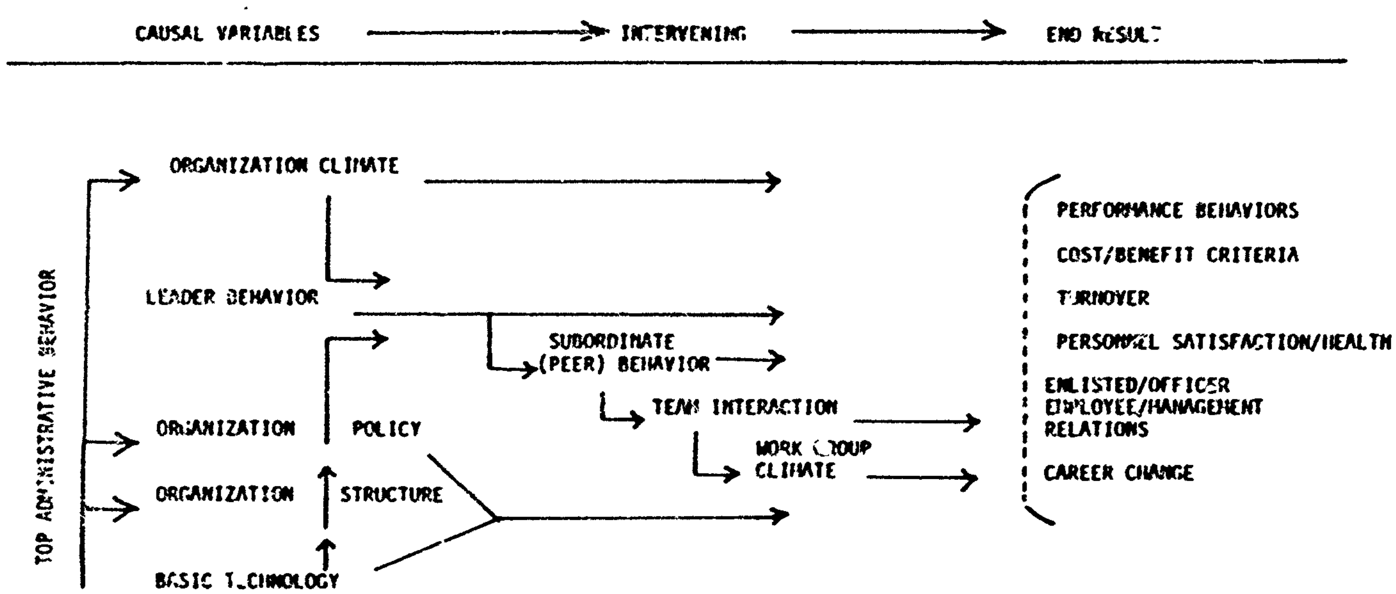


Figure 4 A SYSTEMS APPROACH TO MILITARY EFFECTIVENESS

Adapted from RENIS LIKERT HUMAN ORGANIZATION: MANAGEMENT AND VALUE McGraw Hill, 1967

CASUAL VARIABLES	→	INTERVENING	→	END RESULTS
LEADER BEHAVIOR →	PEER BEHAVIOR →	TEAM INTERACTION →	WORK GROUP CLIMATE →	PERFORMANCE
SUPPORT	PEER SUPPORT		TRUST & CONFIDENCE	
GOAL EMPHASIS	GOAL EMPHASIS	GOAL COMMITMENT	MOTIVATION	MILITARY EFFECTIVENESS
TEAM BUILDING	PEER TEAM BUILDING	TEAM COOPERATION	GOAL CLARITY	COST/BENEFIT RESULTS
HELP WITH WORK	PEER HELP WITH WORK	INFLUENCE SHARING	TEAM COLLABORATION	STAFF SATISFACTION
DECISION-MAKING		DECISION PROCESS	DECISION MAKING	ENLISTED/OFFICER RELATIONS
RECEPTIVITY TO SUBORDINATES' IDEAS		COMMUNICATIONS	COMMUNICATIONS	TURNOVER, ABSENTEEISM, ETC.

Figure 5 FACTORS IN EFFECTIVE UNIT LEADERSHIP

Adapted from RENSIS LIKERT HUMAN ORGANIZATION: MANAGEMENT AND VALUE McGraw Hill, 1967

A description of the format for the applied training cirriculum follows; the applied training involves about 30 participants per session, with at least two professional staff. A training input consists of eight two and one-half day periods scheduled over a three month time span. Two periods are described as a phase. A topical outline of the four phases is attached in Appendix A. The four phases are:

I. Focus on Management of Self
II. Focus on the Organization: Management Skills
III. Focus on the Organization - Supervisory Skills
IV. Management in the Military System

The sessions would provide the input of information, theory and substantive content. Additionally, the sessions include experiential exercises, and role play simulations to try out new alternative approaches, plus opportunities to practice new behaviors in preparation for application in the real world on-the-job.

The process approach is carefully designed to meet the needs of the participants and training sessions are adapted to participants' requirements and pace. Presentations will be supplemented with written manuals that will develop the participants knowledge and awareness of the dynamics of functional management in the military organizational structure. This will be related to successful performance and career productivity in developing appropriate forms, rules and styles of appropriate behavior and communication between women and men in the armed forces.

Specialized emphasis is designed to specifically respond to needs of each special training group. For example, as enlisted personnel, groups 1_T - 4_T would receive emphasis on pre-supervisory and supervisory imput. In contrast assuming officer groups 5_T - 8_T have already been exposed to supervisory patterns, emphasis would be placed on management input. If any officer groups lacked exposure to supervisory patterns, then this would be provided in the special training sessions. In the mixed groups (M*), the initial sessions would be held in separate sex groups, developing understanding of self and others prior to point of developing specific process learnings together. Both male and female participants would then be combined in sessions II, III, and IV.

The development needs of the trainees are addressed in a common core of content in the general sessions. Each session is designed for full participation of the trainees. Three subsequent aspects address the developmental needs of each person separately. (1) at the end of each training segment a brief period is provided for the individual trainee to identify and associate the specific meaning for her/his development and plan for try out in the interim. (2) Additionally, there are scheduled individual "briefing" periods for each trainee to discuss her/his own developmental needs and issues and to receive specific guidance and counseling from the staff. (3) Opportunities for career planning are provided outside of the training assignment to those who wish to focus long range plans,

In summary, it would appear that analytic approaches to optimization in adaptive instruction are possible, and that the treatment intervention broadens the choice in method objectives considering the many faceted aspects of the problem, that the research addresses. Although the technical problems in implementing the research design at three sites are formidable, they do not appear intractable.

RESEARCH ACTIVITY: Evaluation Effectiveness of Applied Training

The research activity proposed for this project is the use of a computer-managed value rating (CMVR) model for evaluating the effectiveness of an applied training program. The model derives from a system developed by Collet (1972) to compare outcomes of simulated evaluation projects with the addition of value rating procedures adapted from Edwards (1977) before presenting the CMVR evaluation model in its entirety, it is appropriate to describe the measurement technique which is its salient characteristic.

It is the aim of the CMVR model to express all evaluation results (outcome variables) in a simple metric scale (Labeled E), which is intuitively meaningful to laymen, yet amenable to statistical manipulation. In addition, it is expected that the proposed E scales will provide both a means of estimating the cost of total achievement of program objectives, and a direct comparison of the cost-effectiveness of programs having fundamentally different objectives.

E-SCALE To illustrate the basic concepts of the model, assume that there are pre and post scores on several standardized tests for each person. Computer Managed Value Voting refers to a procedure in which a computer interacts with a relevant set of experts to develop, for each dependent variable (standardized test), a formula for transforming each raw score into an effectiveness index or E score, where E is a number ranging from 0.00 to 1.00 which represents the degree to which the objective as measured by that variable has been achieved. The basic function of the CMVR computer program is first, to help each expert validly represent his/her values in metric form, and second, to achieve an optimal degree of consensus among experts. The computer output consists of a set of transformation rules. The specific strategies for developing these transformation rules and suggestions for assessing measurement error are discussed under the heading of research strategies.

The description assumes that objectives are stated in terms of absolute achievement, that is, referenced to specific criterion levels for success and non-success. (The procedure adapted to studies of objectives stated in relative terms.)

COMBINING DEPENDENT VARIABLES: After the transformation operations, both the pre and post-treatment performance of each person is represented by a set of E scores. To get a single estimate of a person's overall performance, it is necessary to determine the relative importance of each dependent variable in assessing the achievement of the program objective. The CMVR computer program can use essentially the same strategy as for the development of E scores to interact with the experts to produce a set of normalized weights (i.e. the weights sum to 1.00) which possess this property. The sum of the cross-products of a person's E scores and the associated weight would then produce a composite E score for both pre and post tests. These composite E's would be numbers ranging for 0.00 to 1.00 representing the degree of achievement of the objective as measured by the ENTIRE SET of dependent variables.

MULTIPLE OBJECTIVES: In the above illustration a single program objective was assumed for the sake of simplicity. However, the procedure is easily adaptable to the assessment of programs having multiple objectives by using the above procedure to develop a set of weights representing the relative importance of each objective in achieving a successful program. Strategies for dealing with various levels of objectives (e.g., terminal vs. enabling objectives) are suggested in a subsequent section.

PROGRAM EFFECTIVENESS: One of the convenient attributes of the E metric is that the mean E score (both for the individual dependent variables and for the composite) represents the degree to which the entire group achieved the objectives. The difference between pre and post composite E means (mean E gain) would be a measure of the degree of movement towards the objectives during the administration of the program.

ESTIMATING PROGRAM COST: The mean E gain should be useful in developing generalizable estimates of the program cost. The actual cost of the program divided by the total E gain would provide an estimate of the unit cost of moving one person from complete non-achievement to complete achievement of objectives, assuming uniform growth. This UNIT COST ought to be useful in predicting the cost of moving a new group from their initial performance level to the desire criterion.

ASPECTS OF THE EVALUATION TECHNOLOGY [1]

- Program analysis
- Taxonomy of Evidence
- Concept of lexicographic processing

[1] Detailed explanation and derivation of each is available on request by writing Marilyn E. Harris, Ph.D, 806 Metropolitan Bldg., Flint, Mi 48502

- Concept of theoretical evidence
- Measurement techniques:
  - E transformation
  - importance weights
  - error assessment in E indices
- Simplication through computer management

An important first step in evaluation is to analyze the relationship between instructional activities and intended outcome. The power of this procedure is that it highlights potential weak links. Information about the efficacy of each instructional transaction is important for diagnostic purpose in overall goal.

An example, Figure 3, defines the task to evaluate a program to train military personnel in effective knowledge and skills with respect to its terminal objective of improving military effectiveness. The inferential chain required to link this instructional activity to the designed outcome behavior is illustrated in Figure 3. Each of the numbered boxes represents a potential source of information. If an evaluation project's only interest is in determining whether or not the terminal objective was achieved, information about personnel achievement would provide sufficient evidence for the decision. However, if the purpose is to provide either formative evaluation information or information about the degree to which the terminal objective had been achieved, observations should be made at each of the points between 1 and 17. This would be particularly true in a longterm project such as this example, since one would expect the program effects to slowly filter further and further down the inferential chain with the passage of time.

PROJECTED DELIVERABLES

The following list is intended only to suggest the character and extent of the potential products:

- Comprehensive description of the applied training program with manual
- Evidence of the degree to which the program was actually implemented
- Evidence of the degree to which the program achieved its objectives
- A description of the CMVR evaluation model implementation
- Evidence of the utility of the CMVR model

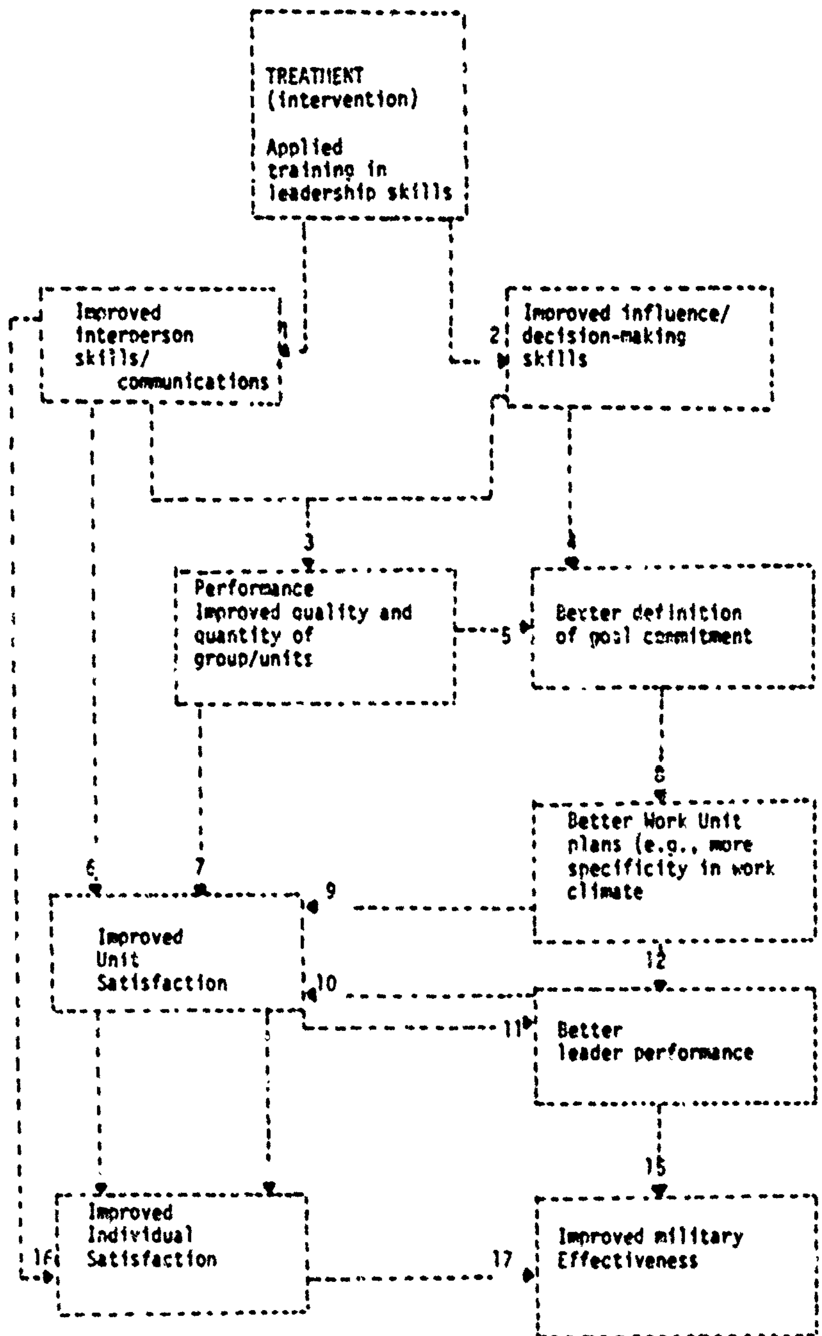


FIGURE 3: Sample inferential chain in an instructional activity

- Implications for implementing the training program and CMVR techniques at other installations
- Manual explicating the CMVR evaluation techniques. This would include practical examples and specific user instructions
- Computer programs and other technical products associated with the CMVR evaluation technique. All programs would be transportable to computer installations with FORTRAN IV capability. (Memory requirements are anticipated to be minimal.)

The expected end-results are both tangible and intangible. In private sector applications, intangible results are documented in reduced stress and tension on-the-job, attractiveness to non monetary factors, positive attitudes and increased trust and team spirit. Tangible results are evidence for example: by the acquisition of increased leadership skills in decision making and problem solving, improved communication skills, interpersonal and organizational and handling of conflict in on-the-job settings. Further, participants acquire basic management skills in planning, organizing, staffing, directing and controlling. These improvements should produce more effective performance of the unit in which the women trainees are assigned, including less lost time and better use of total personnel resources.

The tested benefits to each trainee are:

- Identified level of competence in specific areas
- Improved and more confident self-image
- Identified goal directed behaviors
- Identified "how to process skills immediately applicable both on the job and off the job
- Measured performance effectiveness
- Acquired technical skills applicable to both traditional non-traditional job processes
- Experienced career counseling and job enrichment planning
- Organizational understanding and prespective

In the private sector, these improvement have typically resulted in a minimal budget savings of 5% (range 5% - 20%) for the organization. Similar results are anticipated in a military application of these techniques.

In a standard private sector application, a "training of trainees" program is developed to train the trainees (i.e., the people in the organization who will conduct subsequent training sessions). A similar approach would be expected to be employed in the military. Thus, a turn-key program is envisioned.

Additionally, at the conclusion of the applied training program it will be possible to identify:

- The consequences of integrating significant numbers of women into career positions (e.g., positions having career ladders) and identify career ladder implications in training adoption
- Recommended percentage of integration to give best cost effectiveness results
- The percentage of personnel (male and female) benefited by the applied training; and specific benefits through objectives successfully met
- The degree of benefit to the personnel
- Attractiveness of the job/career to individual
- The benefits to the organization
- The cost savings in human resource utilization qualified performance/unit cost (dollar figures)
- Attitude change from confronted sex differences
- The durability of the training over time
- Identification of key factors in military effectivess
- The social acceptability of the program in the military
- More attractiveness of non-monetary aspects of military life.

CONCLUSION

Considerable study and review of the several social and military factors which affect military effectiveness reveals that the armed forces have similar problems that exist in the private sector as related to the integration of women in the work world. The areas where the unknowns are, are also very similar even including the combat issue, when effectiveness is focussed as the measurable dynamic. In the private sector, large firms are experimenting with similar techniques as those in the applied training approach throughout whole systems, and finding that change is possible.

Empirical conclusions recognize the changes must be a planned change intentional and of one's own volitions; in contrast to emergent or revolutionary change. The planned change must be heavily oriented toward education of all personnel -- education related to the social psychological aspects of human interaction both male and female interacting in the work setting. Understandings of the socialization processes and its consequences provides a real basis for understanding sex differences and to value individual differences over sex differences -- which leads to selection of personnel based on the individual's own ability. Effectiveness clearly will be affected as we gain skill in identifying the components of a task, how to train for it and how to assess successful performance. Then as competence criteria are met in a population effectiveness maybe attained.

Cost effectiveness is likely to result in the integration of women into the military as well as in the private sector as long as women clearly remain in the minority and are trying to demonstrate equality. The amount of savings will be dependent on the quality of the volunteers recruited.

In order to construct useable theory in the area of female integration in the armed forces we believe the applied training approach is worth testing in a comprehensive research project as described here. It is important for the military institution to take initiative in providing the educational intervention model to confront change of the salient dependent variable in a comprehensive research-study approach. And while the applied training and its evaluation is still far from a finished product, it offers one way of resolving a number of fundamental problems in military effectiveness.

APPENDIX A - TOPICAL OUTLINE OF PHASES

PHASE I: FOCUS ON MANAGEMENT OF SELF

Opportunities to understand, discuss and experience the dynamics of the search for identity in a society of many-changing roles.

- Personal issues: male/female needs for approval and success, feelings of ambivalence, conflict, avoidance, sexuality
- Stereotypes and sex roles: myths, facts, reality, socialization and change
- Motivation and achievement: socialization, drives, needs
- Legal boundaries and constraints: laws, processes and procedures, support and action
- Values clarification: value basis, development and ethics
- Decision making and action: data collection, analysis, effects and consequences, evaluation
- Interpersonal communications: cycle, openness and constructive feedback, language

Integrating an image--developing an assertive, affirmative you.

PHASE II: FOCUS ON THE ORGANIZATION: LEADERSHIP ATTITUDES

Opportunities to learn, sharpen skills and try out new methods for productive performance in job and work settings.

- Organizational communications skills: use of information, systematic flow, written vs. oral
- Decision making and problem solving: diagnosis
- Leadership: functional, resource utilization, variability in style, task and maintenance roles

* Goal setting and action planning: specification, focus, diagnosis direction, designing, responsibility

* Intervention and use of influence: alternatives, tasks, types, practice

* Conflict management: awareness, ownership, expression and articulation, alternatives in dealing

* Accountability and responsibility taking: specifying task and time, checking

* Management process: planning, organizing, staffing, direction, controlling

Working in actual situations to develop skills and useful job attitudes for use outside the workshop setting.

PHASE III: FOCUS ON THE ORGANIZATION: SUPERVISORY SKILLS

Opportunities to develop specific skills applicable on the job both traditional and non-traditional.

* Teambuilding: resource identification, relevant goals, integration of affect

* Delegation: identifying specifics, acting, growing responsibility and authority

* Time Management: planning, implementation, discipline

* Performance Evaluation: values, improvement, assessment

* Training others: psychology, theory, practice, experiences

* Management Development - Goal setting, long term development

Integrating skills in on-the-job situations considering planned change.

PHASE IV: MANAGEMENT IN THE MILITARY SYSTEM

Opportunities to plan specifically for your future needs and the needs of others. Learn how to use new knowledge and skills in the military.

- Collaboration and the helping process: inquiry, needs, values, problem identification
- Planning for the future: long range planning, phasing, development and implementation
- Support groups: components, need, criteria and use
- Transition and the changing roles: resistance, support, the change process
- Mobility: role specification, allocation, and programming, organizational development

Planning and interacting with others to develop and test the reality of methods learned in the system.

REFERENCES

Bass, B.M., Krusell, J., & Alexander, R.A. Male managers' attitudes toward working women. American Behavioral Scientist, 1971,15,221-236.

Bond, J.R. & Binacke, W. E. Coalitions in mixed sex triads. Sociometry, 1961, 24, 61-75.

Binkin, Martin and Bach, Shirley J. Women and the Military, The Brookings Institute, 1977

Collet, LeVerne S. and others. "Formative and Summative Evaluation of the FEHR-Practicum Training Model." Final Report- U.S.H.E.W. Project No. 2-0035 June, 1975.

Day, D.R. & Stogdill, R.M. Leader behavior of male and female supervisors: A comparative study. Personnel Psychology, 1972, 25, 353-360.

Edwards, Ward. "How to use Multi-attribute Utility Measurement for Social Decision Making," TEME TRANSACTIONS AND SYSTEMS, MAN AND CYBERNETICS. Vol. SMC-7 #5, May, 1977.

Fidell, L.S. Empirical verification of sex discrimination in hiring practices in psychology. American Psychologist, 1970, 25, 1094-1098.

Kootz, E.D. Women's bureau looks to the future. Monthly Labor Review, 1970, 93, 3-9.

Maier, N.R. Male versus female discussion leaders. Personnel Psychology, 1970, 23, 355-361

Marine, Gene, A Male Guide to Women's Liberation, Avon Books, New York, 1972.

Mathews, E. Employment implications of psychological characteristics of men and women. In W. E Katzel, & W. C. Byham (Eds.), Women in the Work Force. New York: Behavioral Publications, 1972.

Megargee, E. I. Influence of sex roles on the manifestation of leadership. Journal of Applied Psychology, 1969, 53, 377-382.

Scandura, J.M. Problem Solving: A Structural/Process Approach (with instructional implications) New York: Academic Press, in press.

U.S. Department of Labor. Women's Bureau. Fact sheet on the earnins gap. Washington, D.C., 1971.

Likert, Rensis. The Human Organization: Its management and value, McGraw Hill, 1967.

Funktionsanalyse Personalstruktur+)

Presentation

given by Rear Admiral G. Fiebig / Colonel H.E. Seuberlich
on the occasion of the 19th Conference of the Military
Testing Association at San Antonio, Texas, 1977

I. Introduction

(Slide 1)

The "Funktionsanalyse Personalstruktur"+) which will be dealt with here today constitutes a part of the fundamental work aiming at a new manpower structure of the armed forces of the Federal Republic of Germany. This fundamental work is based on the results of the investigations carried out by the Manpower Structure Commission of the Federal Ministry of Defense in 1971; the work of that Commission was expounded before the MTA already on the occasion of the annual conferences of 1972 and 1973. Nevertheless I would like to outline, as an introduction, a few fundamental deliberations of the Manpower Structure Commission as well as the essential elements of their proposals, in order to delineate once more the interrelations and dependencies of that work with regard to the instruments of the "Funktionsanalyse Personalstruktur"+).

When the causes of the continuously increasing difficulties in our armed forces with regard to personnel were investigated it was realized at that time that the problems with which the Federal Armed Forces are faced in the personnel sector cannot be solved at short term. Only a thoroughly new manpower structure of the armed forces,

+) Analysis of the functions pertaining to a job

which corresponds to the industrial society with its principle of division of labour, would be able to bring about a decisive improvement. The main problem seemed to be the following:

In order to be able to maintain their effect of deterrence, the armed forces need continually new and more and more efficient weapon systems which, however, require a steadily increasing proficiency of the personnel.

(Slide 2)

For this reason the systematic build-up of a modern manpower structure is indispensable. That structure must:

1. correspond to the aim of the organization;

2. take into account the capabilities, the age, the personal characteristics and the professional expectations of the individual soldier, and

3. be adapted as far as possible to the external conditions, such as politico-economic developments and labor market.

The basis of the further deliberations of the Commission was that the entire service law has its original point of reference in the individual functions. Under this aspect, an essential element of the concept was the proposal systematically to record and arrange in a new way the interrelations between function, responsibility, pay and rank. Such a reorganization could only be attained on the basis of the description, analysis and assessment of all functions in the armed forces.

(Slide 3)

This reorganization shall inter alia serve for solving the following problems:

- determination of a system-oriented personnel organization

- indication of job characteristics

- development of correlated profiles of requirements which must be fulfilled and of correlated profiles of qualification

- building up of assignment packages with employment-oriented training, and

- weighting of functions in stages.

For that purpose it is necessary to consider the individual functions with regard to the correlation between performance and efficiency. That means that those jobs which are interrelated with regard to the work to be performed and in which the cooperation is directed towards a specific air or result shall be combined. In other words: the achievements obtained by the work on several jobs may be necessary for attaining a specific aim.

(Slide 4)

Such a correlation between performance and efficiency is characteristic of units and sub-units the organization of which is based on the principle of division of labor. This point is particularly important because no foundations for that work in the sense of the goals pursued by the Commission were available, and therefore it became necessary to develop specific instruments.

II. Other Projects

Apart from the Manpower Structure Commission, also other commissions were working at the same time in the domain of the Federal Minister of Defense.

(Slide 5)

In the same year, 1971, reports were submitted by:

- the Force Structure Commission (its first report on equity in induction)
- the Education Commission (the expert report on the reorganization of education and training).

It had a negative effect that the Federal Government ordered immediate measures to be taken in the field of training and education, for example by the foundation of universities of the Federal Armed Forces, although a reorganization of the armed forces ought to have been initiated only after submission of the second report of the Force Structure Commission containing proposals for a new force structure at the end of 1972. For this reason it was not possible in each case to coordinate the interdependencies in a logical manner appropriate to the matter in question. This also impeded the fundamental work in the field of manpower structure sometimes considerably. But the overall gains remained unchanged.

The work aiming at the development of instruments for the description, analysis and assessment of functions in the armed forces began as early as in 1971. At first, certain procedures applied in economy, in public service and in friendly forces were investigated, but these procedures did not correspond to the goals of the Manpower Structure Commission. Thus, the Federal Armed Forces were to some extent faced

with new ground in the scientific field, because all functions to be performed in compliance with the mission assigned to the armed forces had to be recorded systematically and according to uniform aspects. For that purpose, it was on the one hand necessary to subdivide functions into specific elements, and on the other hand to assign skills, knowledge and capabilities required for a function (a job) to the appropriate function elements. Thus a two-stage set of instruments, called "Funktionsanalyse Personalstruktur"+), was created with the support of scientists.

(Slide 6)

In 1973 this set of instruments called "Funktionsanalyse Personalstruktur"+) was subjected to a first comprehensive testing. 3,400 soldiers of all ranks were interrogated according to the principle of qualitative random selection; the soldiers concerned were engaged in the following fields:

- infantry employment,
- aeronautical employment,
- nautical employment,
- personnel administration.

The evaluation as a whole had a satisfactory result. But the horizon of expectations of the Federal Ministry of Defense was not yet entirely satisfied.

(Slide 7)

+) Analysis of the functions pertaining to a job

The main deficiencies were:

- collection of too comprehensive data material,
- neglect of skills as compared with knowledge,
- lack of a scale for determining levels of objectives of learning.

For this reason the set of instruments "Funktionsanalyse Personalstruktur"+) was systematically revised from 1974 onward. Points of main effort were the development of appropriate ADP program structures as well as a repeatedly extended scientific support of the project, which concerned specific items of the latter.

(Slide 8)

You can see the results in this schematic representation of the

FUNKTIONSANALYSE PERSONALSTRUKTUR+)

The flow diagram is divided into seven phases. The three main elements are

- the determination and registration of tasks, in phases 1 to 3
- the analysis of requirements which must be fulfilled, in phases 4 to 6
- the summarizing evaluation, in phase 7.

+) Analysis of the functions pertaining to a job

At present, all individual measures are based on this reliable procedure. Thus it was possible to reduce from 1976 onward the scientific research work. Since that time, the work of the scientists gradually assumes the character of project assistance, as the work can be done to an ever increasing extent by the Federal Armed Forces themselves in the future. I shall revert to this matter when Colonel Seuberlich will have presented to you the instruments in detail.

III. The Inventory of Instruments

(Slide 8)

The multi-stage inventory of instruments called "Funktionsanalyse Personalstruktur"+) serves for the uniform description and analysis of all functions to be performed in the armed forces and for the collection of evaluation criteria in 7 phases and 35 steps.

(Slide 9)

"Function" is to be understood as the totality of tasks to be performed in a particular job. Thus the task is the main element of description to which the requirements which must be fulfilled are referred.

The prerequisite for comparability within the armed forces is a uniform description of tasks. The individual description elements and criteria are to be formulated as uniformly as possible and as specifically as necessary for all functions in all services.

By "task" we understand the totality of purpose-oriented actions which are in direct correlation with regard to performance and

---

+) Analysis of the functions pertaining to a job

efficiency. Within the meaning of the "Funktionsanalyse Personalstruktur"+) the tasks and the relevant definitions of specific forms of tasks are not only elements of a function, but also constitute the basis for the distribution of responsibilities, and thus for the classification of the function concerned into an organisational structure.

We distinguish between five groups of tasks:

1. <u>Tasks of tactical command and control</u>

   The aim of the task "tactical command and control" is to "combine, to organize and to move means of combat in an appropriate manner, taking into account the situation, and to bring them into interaction in combat".
   These tasks imply above all a directing responsibility and may become effective with soldiers down to NCO.

2. <u>Military specialist tasks</u>

   These are tasks listed in the "Catalogue of Functions of the Federal Armed Forces" and subdivided into approximately 60 specialized groups.

3. <u>General tasks of superiors</u>

   These are tasks resulting from the appointment in question or from the capacity as a superior, pursuant to the directive governing military superior-subordinate relations, but not from a specific military function.

4. <u>General tasks</u>

   These are tasks pertaining to another specialized group which do not belong directly to the proper specialist tasks to be performed in a specific specialist function.

+) Analysis of functions pertaining to a job

5. Additional task/function

Task/function performed temporarily or permanently by soldiers during the time they belong to the unit/agency in question, beside their military specialist tasks, their general tasks as superiors and their general tasks depending on the mission concerned and on the requirements in the unit/agency. Such a task/function is not bound to the proper tasks pertaining to a job.

We distinguish, with regard to the inventory of instruments, also specific forms of tasks - apart from these groups of tasks. By specific forms of tasks we understand particular parts of tasks. As a rule, they comprise several actions all of which are characterized by the same degree of executing responsibility or directing responsibility and are referred to procedures or subjects having the same degree of difficulty. Within the meaning of the "Funktionsanalyse Personalstruktur"+), it shall be possible to perform these actions in a single job; but this does not preclude that the holders of several jobs may participate in the actual performance.

I have now explained a few terms which are very important for understanding and using the inventory of instruments of the "Funktionsanalyse Personalstruktur"+), because the aim, a uniform description of tasks, can only be reached if the essential terms are handled in a uniform manner.

I repeat that for the "Funktionsanalyse Personalstruktur"+) the task is the main element of description and that the requirements pertaining to the function in question are referred to the task. The inventory of instruments is subdivided correspondingly and consists of two main complexes:

+) Analysis of the functions pertaining to a job

- Part A of the inventory of instruments serves for the determination and registration of tasks

- Part B serves for the analysis of requirements which must be fulfilled.

The inventory of instruments is used within the scope of this flow diagram consisting of seven phases. As you see, phases 1 to 3 serve for the determination and registration of tasks, up to the job description; phases 4 to 6 comprise the analysis of requirements which must be fulfilled, up to the assignment of capabilities. Phase 7 is scheduled for the summarizing evaluation, up to the setting up of assignment packages, i.e. of those new means for classification in the field of personnel organisation which the Manpower Structure Commission has considered necessary for the purpose of enabling a clearly arranged classification of the new manpower structure of the armed forces.

IV. The flow diagram

Now a few words on the individual phases of the flow diagram (determination and registration of tasks).

(Slide 10)

Phase I

The determination and registration of tasks begins with phase 1, the current development of standardized instruments of inquiry. It concerns lists of tasks, questionnaires on tasks or catalogues of tasks. They are prepared in technical talks with specialist instructors at schools and with competent soldiers experienced in practice. Formulations of tasks, as main elements of description, must fulfil specific requirements. These requirements are summarized in six rules.

(Slide 11)

Rule 1: The task shall be formulated in such a way that the function concerned is expressed in terms showing the goal of the task; it shall also indicate exactly the activity to be performed.

Rule 2: The wording of the tasks shall be unambiguous from a linguistic point of view and formulated in a manner intelligible to all.

Rule 3: A task shall constitute a working cycle/complex clearly delineated with regard to time required and specialty. Several job holders may participate in performing a task.

Rule 4: The formulation of the tasks shall enable a clear delineation.

Rule 5: The tasks shall not be worded in too narrow a way.

Rule 6: The tasks shall not be worded too comprehensively, i.e. they shall not cover the tasks of entire sub-units.

(Slides 12 and 13)

Example (wrong/right)

The various forms of all specific tasks show differences with regard to complexity, executing responsibility and directing responsibility. Correspondingly, one of the five tendency values laid down in ascending order is attributed to the forms of the specific task concerned for each applicable characteristic of the task. I would like to explain these three terms:

(Slide 14)

<u>Complexity</u> is a feature concerning the characterization of contents and the difficulty of tasks. It may be caused by a great multitude and number of subjects and a great variety of equipment and procedures at which or with which the task is performed. But it may also be brought about by the composition and scope of the area of work, conditioned on multiple, far-reaching flow of work as well as on the number of persons participating and by their qualifications.

By <u>executing responsibility</u> we understand the responsibility for purpose-oriented actions, implying the correct application of procedures and methods. The level of the executing responsibility within the scope of a task depends on the one hand on the degree of independence and on the other hand on the significance and the consequences of one's own actions.

By <u>directing responsibility</u> we understand the responsibility for purpose-oriented, guiding and steering influence on the behavior of other men as well as the responsibility for purpose-oriented employment of material in compliance with the task. The level of the directing responsibility within the scope of a task is determined by extent and importance, intensity and consequences of the measures caused with regard to the employment of personnel and material for the purpose of performing the task.

The importance of the individual task is compared with the importance of every other task pertaining to a job and assessed according to its contribution for the fulfilment of the mission of the unit or sub-unit concerned.

(Slide 10)

I would summarize Phase I as follows:

First of all,

- the specialized activities to be investigated are selected and

- selection plans are prepared on the basis of military occupational specialties and of sub-units.

The indications made on the importance of the task as compared with other tasks may differ widely

- "in normal duty"

- as well as "during exercises".

For the "Funktionsanalyse Personalstruktur"+), to perform a task "in normal duty" means to perform it under peacetime conditions or training conditions. For the purpose of differenciation, the designation "during exercises" is used when a task is concerned which is performed under field or simulated war conditions, such as they exist in case of crisis and tension and in a defense emergency.

It goes without saying that also

- "particular strain" caused by

  - time pressure,

  - strenuous attitude during work,

  - work requiring difficult movements,

  - work implying one-sided movements,

+) Analysis of the functions pertaining to a job

- holding of heavy objects, etc.

- as well as the "particular working conditions" caused by

  - temperature,
  - heat radiation,
  - air flow,
  - permanent noise, etc.

  will be registered.

As to the use of the various inquiry instruments, it is extremely important to apply them according to the ADP requirements, because otherwise neither the recording in the various necessary data files nor the evaluation can be ensured.

In the course of the next step, competent soldiers are interrogated according to a given scheme, as to the tasks to be performed by soldiers in the functions and sub-units to be investigated.

During the third step, competent soldiers experienced in practice comment on the pre-formulated tasks.

Thereupon the final specific forms of tasks are laid down on the basis of task characteristics, complexity and responsibility.

These tasks are then recorded by ADP in the data file on tasks and in the lists containing the specific forms of tasks, for the further interrogation on AUTHORIZED tasks, as well as in the catalogues of questionnaires on tasks intended for the interrogation on ACTUALLY performed tasks.

(Slide 15)

Phase II

Phase II comprises an inquiry on ACTUALLY performed tasks and a preliminary inquiry on AUTHORIZED tasks. Selection plans for representative interrogations are prepared for the inquiry on ACTUALLY performed tasks. The soldiers are interrogated in groups of about 30 men. The procedure of answering of questions concerning the catalogues of questionnaires on tasks is simple. The soldiers have to select the indications pertinent to the tasks they perform and to document them with the aid of forms to be marked for documentation by ADP.

The inquiries on AUTHORIZED tasks are carried out in a similar way. In this context it is important to elaborate the concepts on AUTHORIZED tasks, with regard to the distribution of tasks to the various jobs, in the course of talks with competent soldiers according to fixed rules. These soldiers are asked in which jobs which tasks - and which specific forms of tasks - are to be performed. Besides they must indicate which importance is to be attributed to these specific forms of tasks in normal duty and during exercises. Such inquiries on AUTHORIZED tasks can only be carried out by means of carefully articulated questionnaires which are summarized in the lists of specific forms of tasks.

(Slide 16)

Phase III

In this phase, the evaluation of tasks according to the determination and registration of ACTUALLY performed as well as of AUTHORIZED tasks is important. This is done by ADP. The comparison AUTHORIZED/ACTUAL reveals the contrast between the results of the inquiries on

AUTHORIZED and on ACTUALLY performed tasks for a definite domain. It will then be shown once more to the same competent soldiers who have carried out the first assignment of AUTHORIZED tasks in Phase II. Besides, experts in training, in planning with regard to material to be used and in organization will be consulted in Phase III, in order to enable the clarification of divergencies between the results concerning AUTHORIZED and ACTUALLY performed tasks and also the storing of amended AUTHORIZED tasks. The results of that work are uniform job descriptions for the armed forces which will be stored in a data file. Thus these descriptions are already available for various purposes.

(Slide 17, analysis of requirements which must be fulfilled)

<u>Phase IV</u>

The analysis of requirements which must be fulfilled is initiated in this phase. The requirements connected with the tasks to be carried out by a job holder must therefore be determined and recorded during the next steps. The three categories concerned are the following:

- capabilities,
- knowledge,
- skills.

By <u>"capabilities"</u> we understand the relatively long-lasting physical and psychic quality of a person enabling the acquisition or application of skills and knowledge.

By "knowledge" we understand what a person must know for being able to perform a function successfully. Knowledge is acquired primarily by theoretical instruction and is a prerequisite of skills.

By "skills" we understand, on the contrary, the intellectual as well as the physical attainments of a person. Skills are acquired above all during practical training.

Phase IV comprises the preparatory work for the analysis of requirements which must be fulfilled; it begins with the collection of material. It is based on the list of specific forms of tasks and on numerous documents of the Federal Armed Forces, such as training plans, training programs etc. This list must be formulated on a level which is uniform for the Federal Armed Forces with regard to language and contents.

During the next step, the preliminary collections of material are supplemented in talks with specialist instructors and competent soldiers experienced in practice, by adding terms not yet contained in the collections. Besides, the experts correct and check the preliminary collections of material. The collections amended in this way are recorded by ADP and stored in a data file on collections of material.

The concrete preparation of the inquiry begins in the next step. In this context, a balanced relation between instructors at schools and superiors in the troops is particularly important. The selection of the best experts for the assignment of knowledge and skills to the task in question is of decisive importance.

(Slide 18)

Phase V

With the aid of the documents prepared the experts are briefed in this phase in the procedure of assigning knowledge and skills. It is incumbent upon them to assign the necessary objects, fundamentals and procedures to the specific forms of tasks they have to deal with. At the same time they have to establish the required level of knowledge.

Thereupon the experts select, among the assigned procedures, those procedures which, within the scope of execution and of specific forms of tasks, must be mastered as skills. The degree of mastering the skills in question will be determined according to a scale of 5 values.

In the next step, the experts assign, out of the list "Particular strain/working conditions" consisting of about 20 characteristics, those characteristics which are typical in the execution of the specific forms of the tasks concerned. Besides, degrees of strain are assigned to these selected characteristics; the respective scale is divided into three values. These degrees of strain may be referred in different ways to normal duty or to exercises, it is true, but only the highest degree of strain is evaluated.

During the last step, the experts deal with the capabilities required for executing a specific form of a task. They select these capabilities from a catalogue in which 20 capabilities are described. The classification is made according to a scale of five values showing a number of examples. These indications are recorded on an inquiry form capable of being used for ADP.

(Slide 19)

Phase VI

During this phase, the data obtained by the inquiries are checked and stored. The results are printed out in the form of profiles for requirements which must be fulfilled with regard to "specific forms of tasks". Such a profile contains indications concerning:

1. necessary knowledge of

   - objects at which the specific form of the task is executed,

   - objects with which the specific form of the task is executed,

   - fundamentals,

   - procedures and methods applied for executing the specific form of the task;

2. skills required for applying or performing

   - procedures or methods used for executing the specific form of the task;

3. "particular strain / working conditions";

4. capabilities.

The thus established profiles for the requirements which must be fulfilled are submitted to a selected group of experts among specialist instructors and competent soldiers experienced in practice as well as to military psychologists and physicians for occupational medicine for

examination. In cases of evident lack of plausibility the group of experts may revert to indications obtained during phases II and III. The possibilities of correction, e.g. indications on importance and frequency of performance of a task, contained therein may enable amendments in connection with the indications obtained in phase V. The results are processed by ADP and stored in the corresponding data files.

(Slide 20)

Phase VII

During that last phase, the final evaluation and the setting up of assignment packages take place. The results of the determination and registration of tasks and those of the analysis of requirements which must be fulfilled are summarized. The profiles of requirements which must be fulfilled with regard to individual functions of specialized groups within the armed forces are compared. These profiles are classified and arranged in packages after weighting. The descriptions of functions printed out in ADP and the results of cluster analyses are useful for the evaluation and enable the setting up of assignment packages.

The results of this final evaluation procedure are included in the data file for profiles of requirements which must be fulfilled with regard to the functions. The possibilities of application by the requesting agency are more manifold, e.g. the planning of uniform training of the armed forces in compliance with the requirements, the reasonable redistribution of tasks, a steering in personnel matters which is to a large extent adapted to the requirements as well as to the wishes of the individual soldiers.

This is what I wanted to say about the instrument "Funktionsanalyse Personalstruktur"+) and its seven phases.

On the basis of these statements, Admiral Fiebig will give you an outlook on the further development of the "Funktionsanalyse Personalstruktur"+).

V. Outlook

May I sum up these statements by the following valuation:

Planning and organization of the "Funktionsanalyse Personalstruktur"+) have been completed.

(Slide 21)

We have found out that a total of 104 catalogues of questionnaires must be elaborated.

47 of these catalogues cover 50 to 200 military specialist tasks or tasks of superiors, and 57 catalogues up to 50 military specialist tasks. The tasks of tactical command and control are included in these figures.

By the end of 1976, a total of 28 catalogues of questionnaires on tasks and lists of specific forms of tasks, that is about one third of the necessary catalogues, were completed.

By the end of 1977, presumably about two thirds of all catalogues will be elaborated. Then it will be possible to represent about one half of all jobs of the armed forces as well as their functions.

(Slide 22)

+) Analysis of the functions pertaining to a job

For the inquiry on ACTUALLY performed tasks it will be necessary to interrogate a total of approximately 54,000 soldiers. As one team interrogates about 30 soldiers per day, a total of 1,800 interrogation days are required.

This inquiry on ACTUALLY performed tasks shall in the main be completed in 1981.

But the continuous evaluation of intermediate results is possible already now. After completion of the determination and registration of tasks, the available data material will furnish for instance:

(Slide 23)

- task descriptions concerning all kinds of functions in the armed forces, classified according to generic terms, hierarchically arranged pursuant to the degree of difficulty of task performance and to the different degrees of responsibility connected with the task in question, subdivided with regard to directing responsibility and executing responsibility,

- job descriptions, referred to the present organization, with indications on all specific forms of tasks to be executed in the job, and on the distribution of the degrees of importance of the tasks within the job,

- indications on working conditions and environmental conditions as well as on working hours of jobs; particular strain and working conditions referred to tasks,

- indications on danger of accidents/safety of work and on means for ensuring that safety on the jobs,

- indications on the frequency of performance of tasks in normal duty and during exercises,

- indications on licences, permits and evidence that particular regulations have been complied with,

- indications on courses attended,

- indications on appointments,

- indications on performance of general tasks and on the frequency of performance,

- indications on additional functions and on their distribution to jobs,

- indications on contentment with regard to occupation.

The data material is arranged in such a way that it enables uniform evaluations in the armed forces.

(Slides 24 and 25)

This becomes particularly clear when we consider the so-called standard of functions which includes inter alia the following functions:

As these statements show, the use of the instruments of "Funktionsanalyse Personalstruktur"+) has been fully initiated and is energetically supported by all services.

Independently of the organizational regulations, the cooperation of the troops remains after all decisive for the success of the work. Above all it is important to interrogate the right specialist instructors at the schools of the services, for only then it will be possible to determine and register reliably the tasks necessary for performing a function and to describe them according to uniform criteria. Beside the judgment of these theorists at the schools, the interrogation of the practicians in the troops is of course of equal importance. There, soldiers, who know very thoroughly the terms of reference in practice, have to give the relevant information.

Since May 1976, these interrogations have constituted a workload of not insignificant an extent to schools and troops. In this connection, the insight into the importance of this work with regard to a new manpower structure of the armed forces is preponderant over the difficulties which may arise in certain cases. But again and again it becomes evident how important it is to motivate in the right way the persons concerned. This aspect must not be disregarded. It even becomes more and more significant the longer the work lasts.

In the meantime, organization, flow diagram and time schedule have been coordinated in such a way that it may be said that on the whole the term originally pre-planned by the Manpower Structure Commission for the completion of the work concerning a new manpower structure, the year 1981, can in the main be complied with.

+) Analysis of the functions pertaining to a job

The instruments of the analysis of functions which were outlined to you today and which serve for elaborating the new manpower structure of the armed forces of the Federal Republic of Germany constitute an essential prerequisite for the solution of numerous interservice tasks of the Federal Armed Forces. For this reason, the Federal Minister of Defense explicitly ordered in 1976 the continuation of the analysis of functions. But he did so under two premises:

1. The work may be performed only as basic research for the Federal Armed Forces themselves.

2. Conceptual models on the evaluation of jobs within the scope of the analysis of functions must not have any implications on public service as a whole.

This direction by the Federal Minister of Defense was necessary because in our country the Ministry of the Interior - not the Ministry of Defense - is competent for public service.

But independently of our work, the competent department also intensely deals already with these problems, and I would not exclude that fundamental results of our investigations may gain validity for other domains, too.

(Slide 25)

Independently of the aforegoing we see that the analysis of functions will retain its importance also in the far future, for the introduction of new weapon systems and of new command

and control systems as well as organizational changes must again and again be analysed, so that it will be possible to draw conclusions within the Federal Armed Forces, for example with regard to the modification of the manpower structure of the armed forces. For this reason, the instrument "analysis of functions" is built up as an open system which possesses a sufficient flexibility also with regard to future developments and changes of multiple kinds.

SLIDE 1

FUNKTIONSANALYSE PERSONALSTRUKTUR *)

= PART OF THE FUNDAMENTAL WORK

FOR A

NEW MANPOWER STRUCTURE OF THE ARMED FORCES

ON THE BASIS

OF

RESULTS OF INVESTIGATIONS

CARRIED OUT BY THE

MANPOWER STRUCTURE COMMISSION OF THE

FEDERAL MINISTRY OF DEFENSE

IN 1971

*) Analysis of the functions pertaining to a job

SLIDE 2

MODERN MANPOWER STRUCTURE

MUST IN A BALANCED WAY TAKE INTO ACCOUNT:

1. AIM OF ORGANIZATION

2. CAPABILITIES
   PERSONAL CHARACTERISTICS
   PROFESSIONAL EXPECTATIONS } OF THE INDIVIDUAL SOLDIER

3. TECHNICAL
   POLITICO–ECONOMIC
   SOCIAL } DEVELOPMENTS

SLIDE 3

FUNKTIONSANALYSE PERSONALSTRUKTUR *)

SERVES INTER ALIA FOR

1. SYSTEM–ORIENTED PERSONNEL ORGANIZATION

2. PERSONNEL–ORIENTED JOB CHARACTERISTICS

3. ENABLING COMPARISON OF

   PROFILES
   - FOR REQUIREMENTS WHICH MUST BE FULFILLED
   - FOR QUALIFICATIONS

4. BUILDING UP OF ASSIGNMENT PACKAGES

5. WEIGHTING OF FUNCTIONS

*) Analysis of the functions pertaining to a job

SLIDE 4

PROBLEM

RECORDING OF CORRELATION

BETWEEN PERFORMANCE AND EFFICIENCY

IMPORTANT

BECAUSE:

CORRELATION BETWEENPERFORMANCE AND EFFICIENCY

= CHARACTERISTIC OF UNITS AND SUB-UNITS ORGANIZED ACCORDING TO THE PRINCIPLE OF DIVISION OF LABOR

SLIDE 5

IN 1971,

APART FROM THE REPORT OF THE MANPOWER STRUCTORE COMMISSION,

REPORTS WERE ALSO SUBMITTED BY

1. THE FORCE STRUCTORE COMMISSION

   (ON EQUITY IN INDUCTION)

2. THE EDUCATION COMMISSION

   (ON REORGANIZATION OF EDUCATION AND TRAINING)

HEREWITH SOMETHING LIKE A „LESS APPROPRIATE TIMING"

WAS CREATED

SLIDE 6

FUNKTIONSANALYSE PERSONALSTRUKTUR *)

FIRST TESTING IN 1973

WITH 3,400 SOLDIERS

FROM THE FOLLOWING FIELDS / SPECIALIZED GROUPS:

- INFANTRY EMPLOYMENT
- AERONAUTICAL EMPLOYMENT
- NAUTICAL EMPLOYMENT
- PERSONNEL ADMINISTRATION

AT THAT OCCASION

THE FUNCTIONS PERTAINING TO

- 90 MILITARY OCCUPATIONAL SPECIALTIES
OF ALL SERVICES

WERE RECORDED

*) Analysis of the functions pertaining to a job

SLIDE 7

EVALUATION

FIRST TESTING

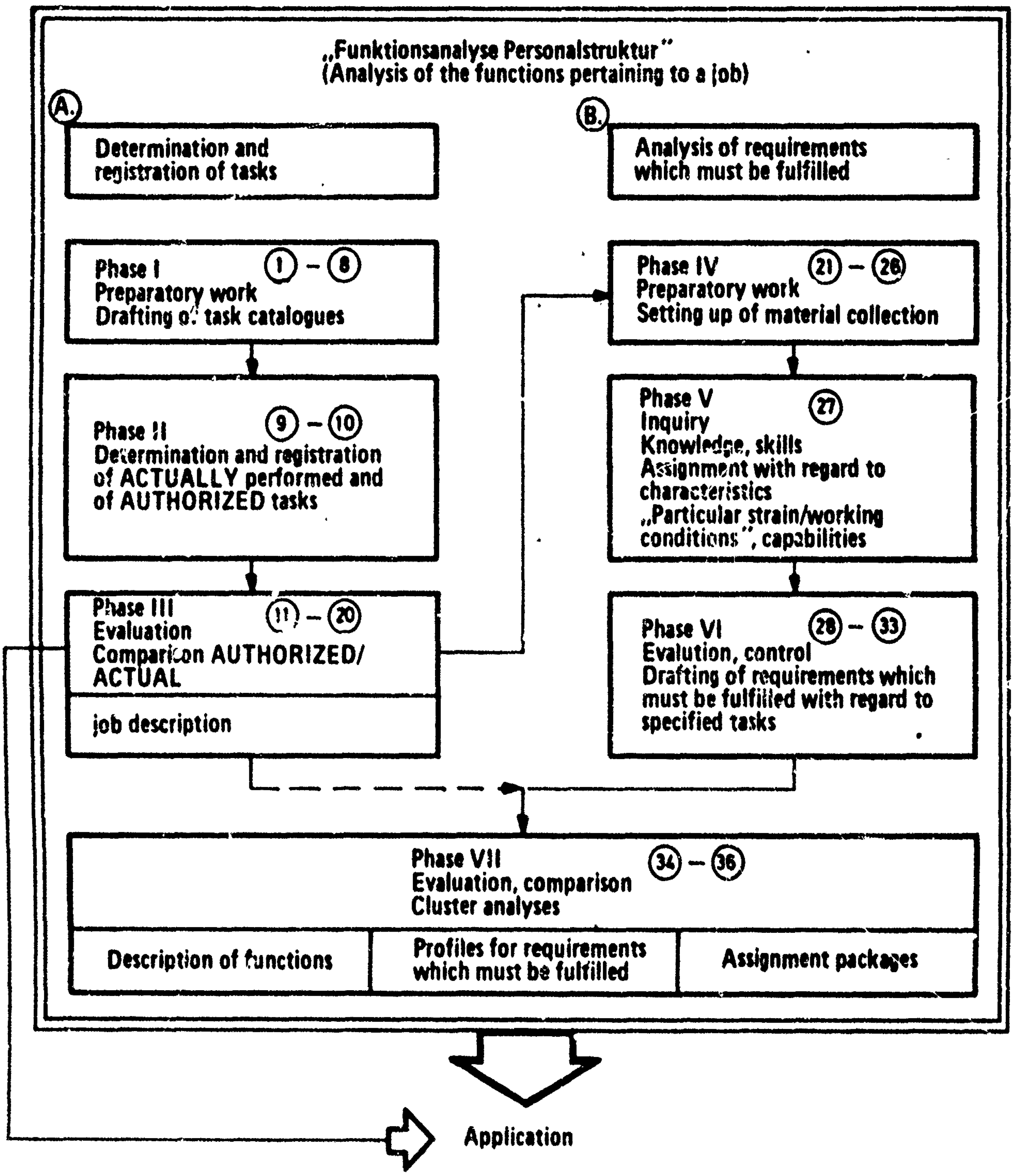
DEFICIENCIES

STILL EXISTING:

- TOO COMPREHENSIVE DATA MATERIAL
- NEGLECT OF CRAFTSMANSHIP
- NO LEADS WITH REGARD TO LEVELS OF OBJECTIVES OF LEARNING

SLIDE 8

Flow Diagram



SLIDE 9

FUNCTION

= TOTALITY OF TASKS (JOB)

TASK

= MAIN ELEMENT OF DESCRIPTION

UNIFORM DESCRIPTION OF TASKS

= PREREQUISITE FOR COMPARABILITY OF FUNCTIONS

SLIDE 9 a

5 GROUPS OF TASKS

1. TASKS OF TACTICAL COMMAND AND CONTROL

2. MILITARY SPECIALIST TASKS

3. GENERAL TASKS OF SUPERIORS

4. GENERAL TASKS

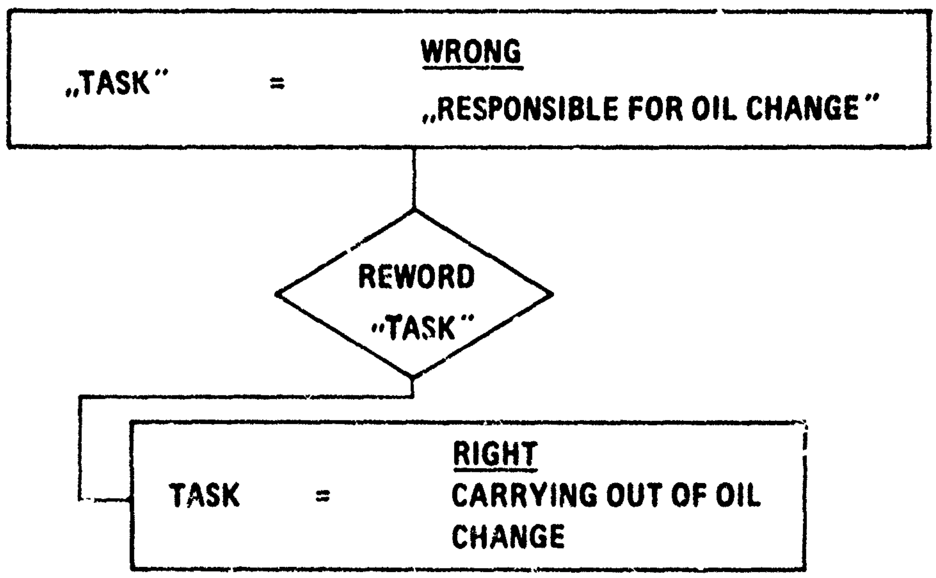
5. ADITIONAL TASKS / FUNCTIONS

SLIDE 11

EXAMPLE CONCERNING RULE 1

AS TO THE FORMULATION OF THE INDIVIDUAL TASKS, THE MODE OF EXPRESSION USED SHOULD BE AS UNIFORM AS POSSIBLE

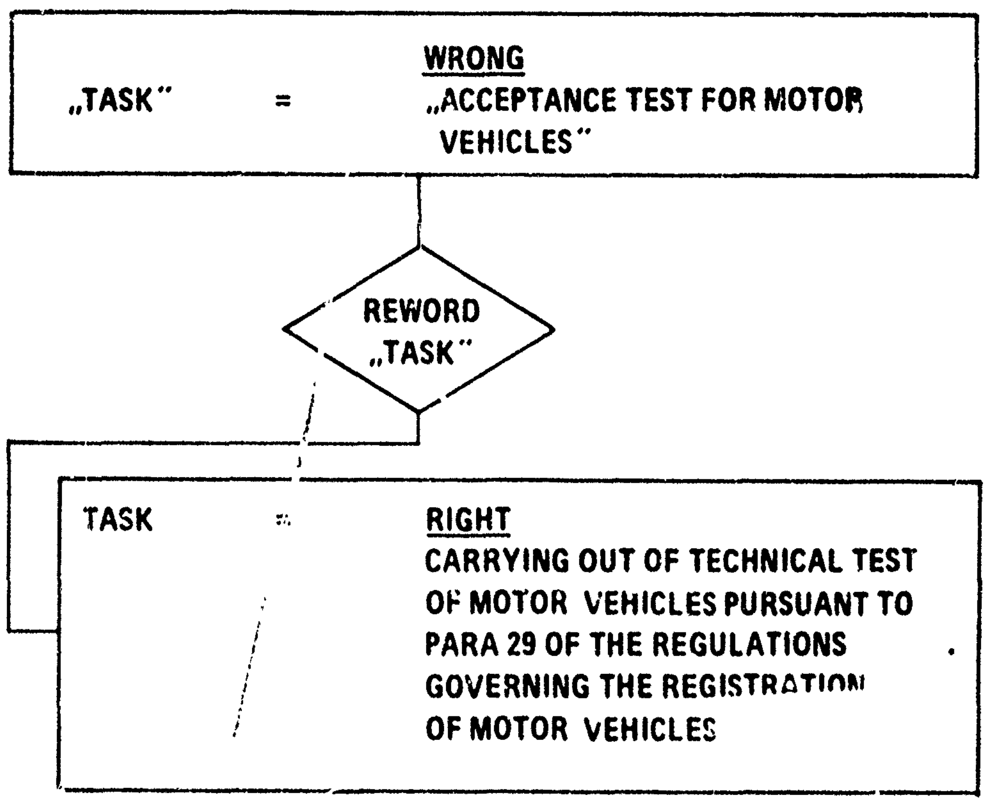
(VERB + PREDICATE)



SLIDE 12

EXAMPLE CONCERNING RULE 2

THE FORMULATIONS OF TASKS CONSTITUTE LEADS FOR THE SERVICEMEN TO BE INTERROGATED WHICH FACILITATE THE ANSWERING OF QUESTIONNAIRES ON TASKS; FOR THIS REASON IT MUST BE AVOIDED THAT THE INTERROGEES FAIL TO ANSWER THE APPLICABLE QUESTIONNAIRE ON TASKS FOR THE ONLY REASON THAT THE FORMULATION OF TASKS IS AMBIGUOUS OR NOT USUAL.

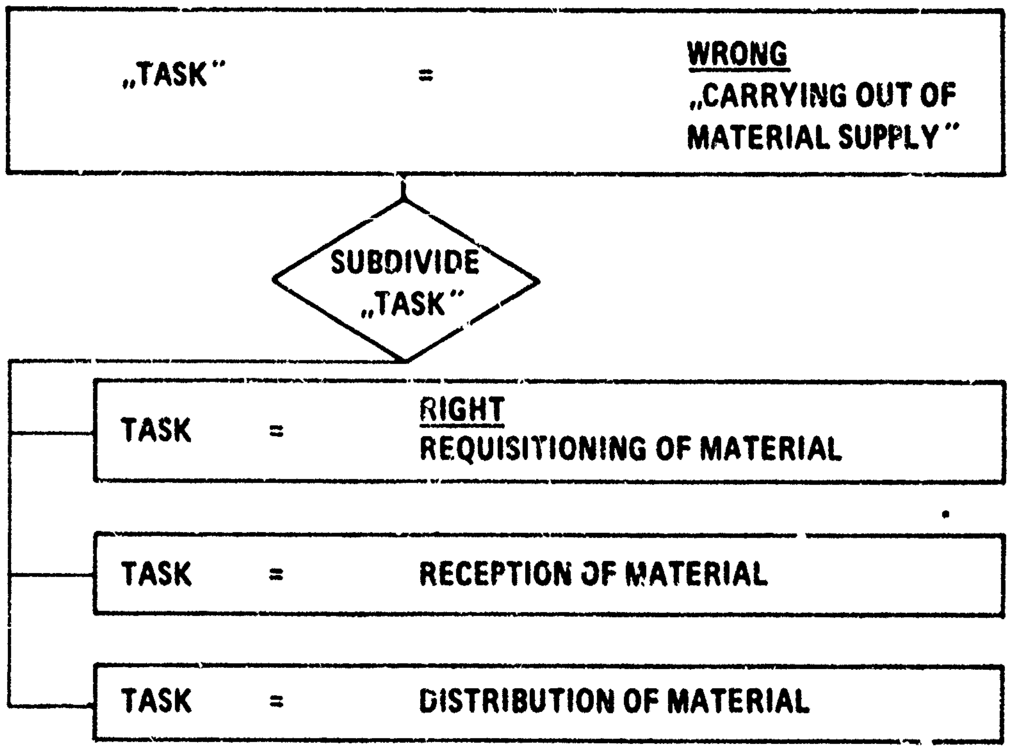


SLIDE 13

EXAMPLE CONCERNING RULE 3

A TASK SHALL CONSTITUTE A WORKING CYCLE / COMPLEX CLEARLY DELINEATED WITH REGARD TO TIME REQUIRED AND SPECIALTY. SEVERAL JOB HOLDERS OF DIFFERENT LEVELS MAY PARTICIPATE IN PERFORMING A TASK.

NOTE: ONE JOB HOLDER MUST BE IN A POSITION TO PERFORM THE SPECIFIC TASK DEFINED; FOR THIS REASON IT MUST NOT CONSTITUTE THE TASK OF A UNIT / SUB-UNIT.



SLIDE 14

COMPLEXITY

= CHARACTERISTIC OF

- CONTENTS
- DEGREE OF DIFFICULTY

} OF A TASK

EXECUTING RESPONSIBILITY

= RESPONSIBILITY FOR ACTING BY APPLICATION OF PROCEDURES OR METHODS

DIRECTING RESPONSIBILITY

= RESPONSIBILITY

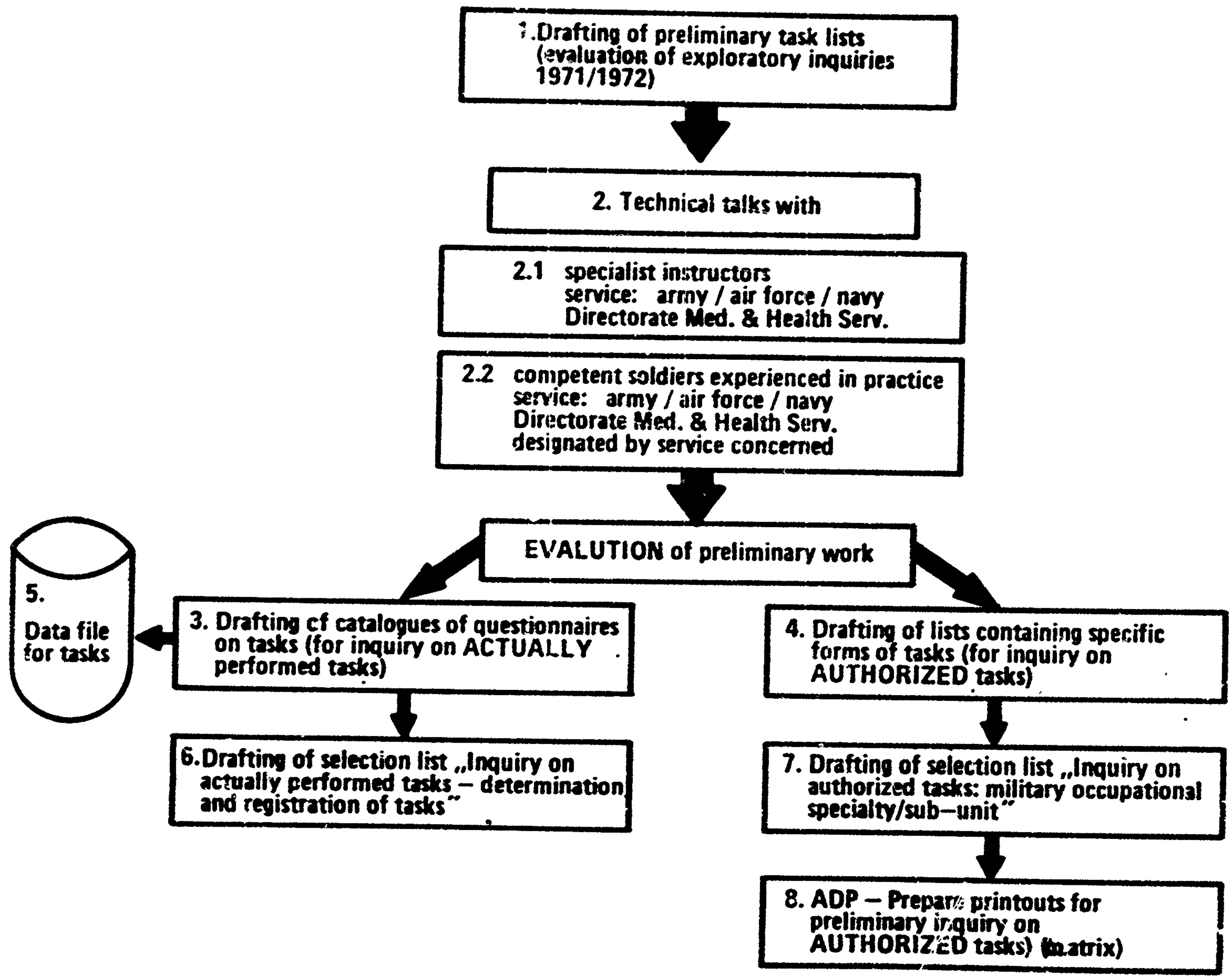
- FOR INFLUENCING BEHAVIOR OF OTHER PEOPLE

  AND

- FOR PURPOSE-ORIENTED USE OF MATERIAL

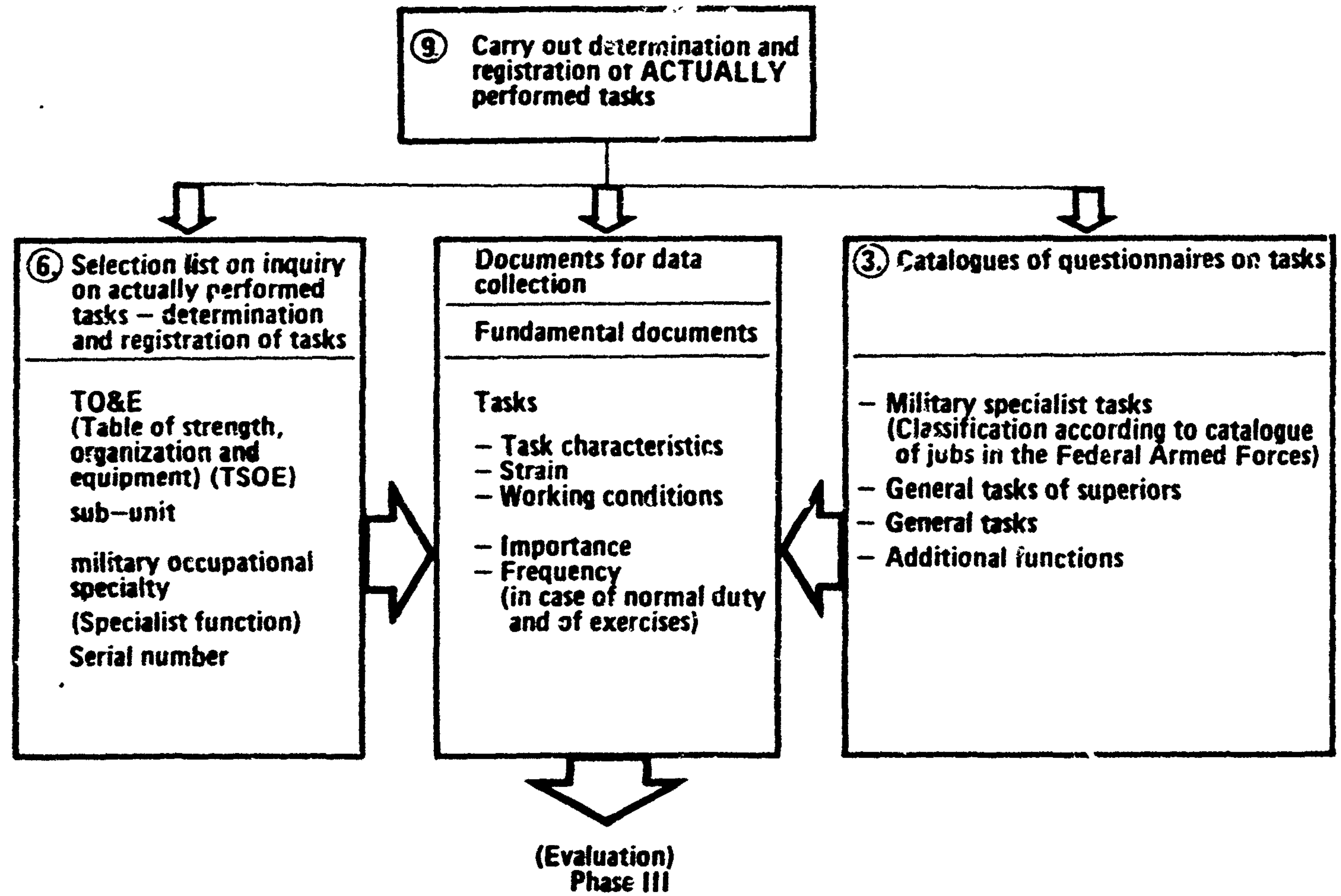
SLIDE 10

Determination and Registration of Tasks Phase I (Preliminary Work)



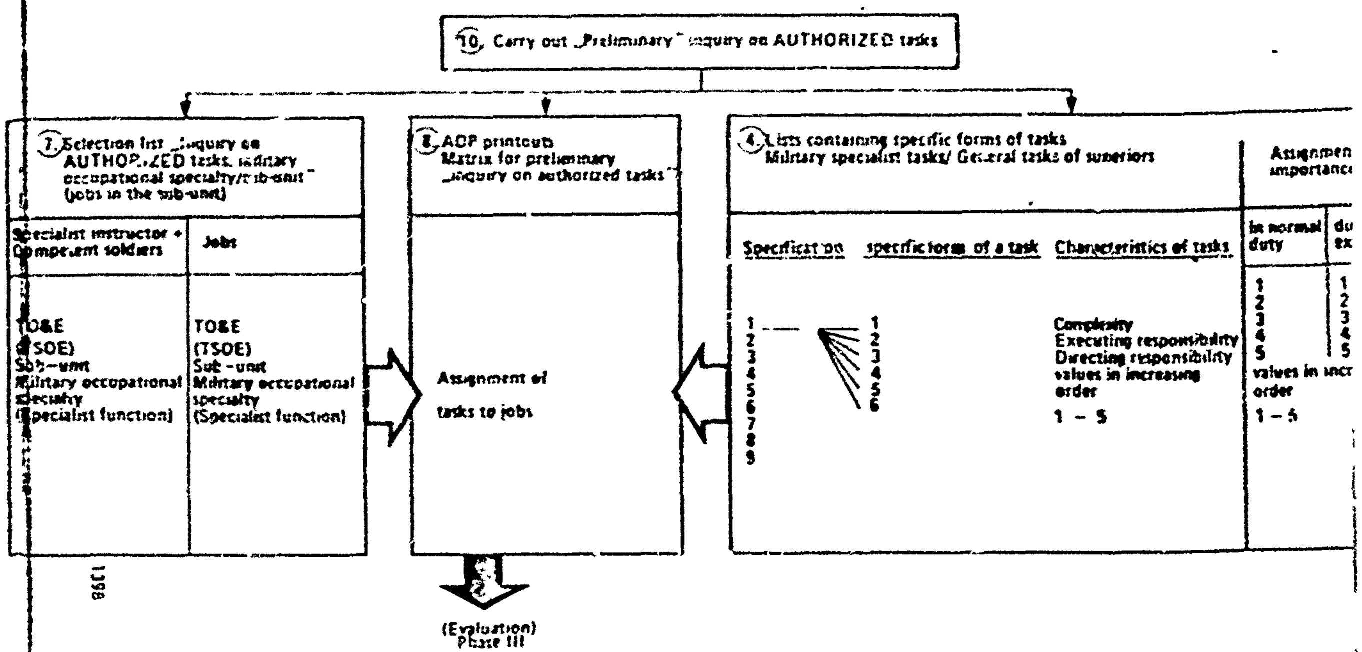
SLIDE 15

Determination and Registration of Tasks Phase II 1 (ACTUAL)



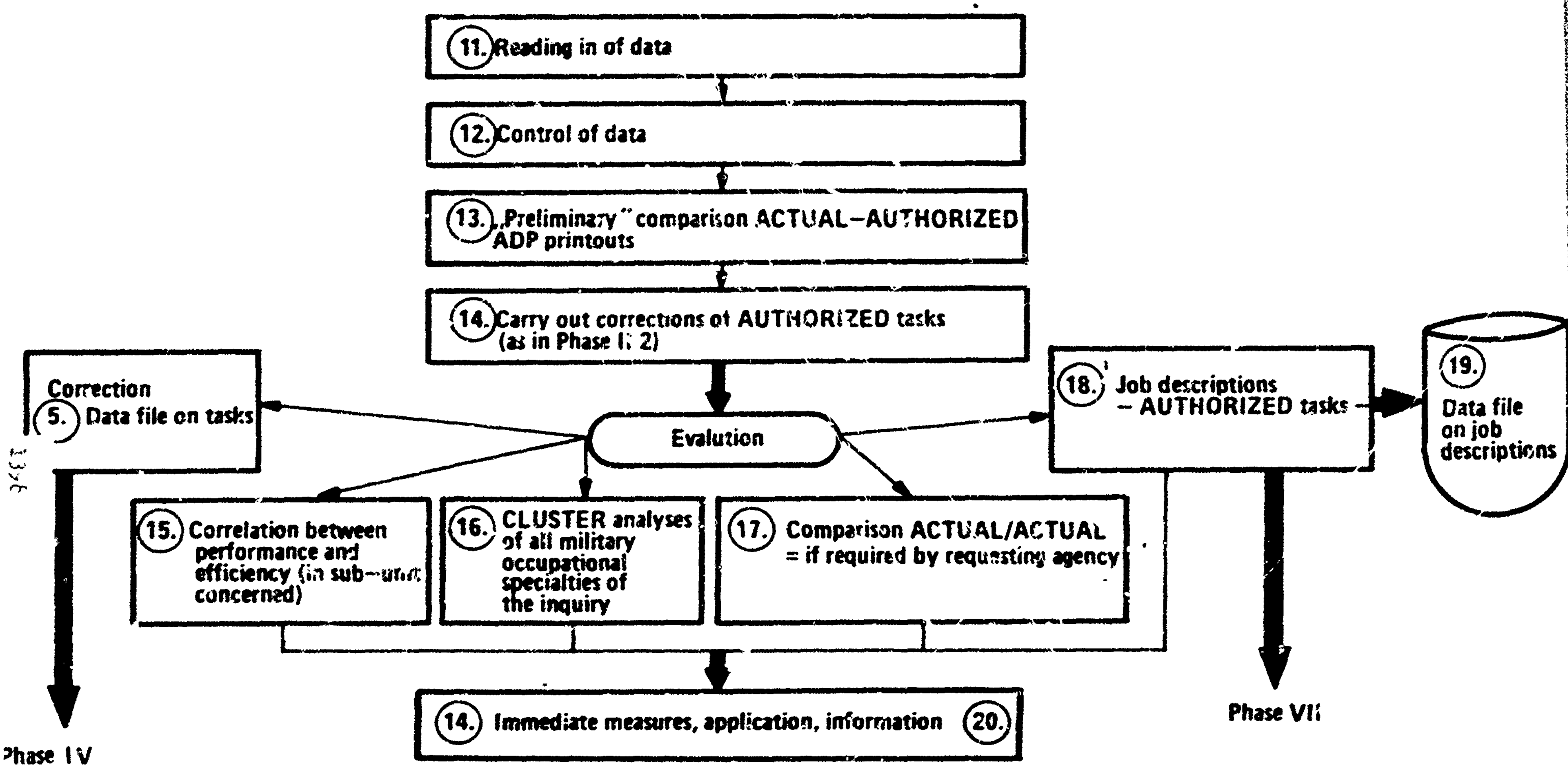
SLIDE 15 a

Determination and Registration of Tasks Phase II 2 (AUTHORIZED)



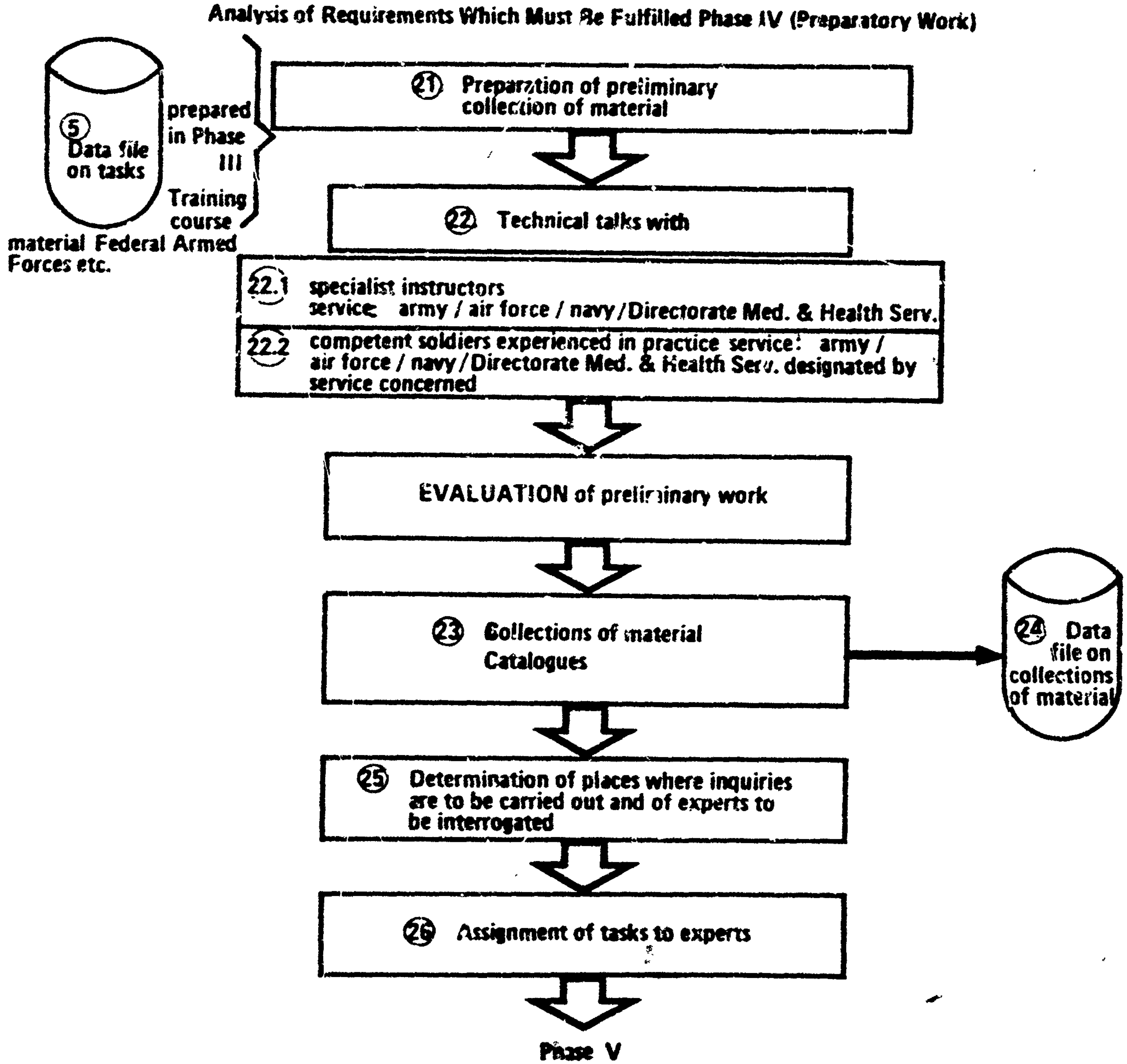
SLIDE 16

Determination and Registration of Tasks Phase III (Evaluation)



1376

SLIDE 17


Analysis of Requirements Which Must Be Fulfilled Phase IV (Preparatory Work)
5 Data file on tasks
prepared in Phase III
Training course
material Federal Armed Forces etc.
21 Preparation of preliminary collection of material
22 Technical talks with
22.1 specialist instructors service: army / air force / navy/Directorate Med. & Health Serv.
22.2 competent soldiers experienced in practice service: army / air force / navy/Directorate Med. & Health Serv. designated by service concerned
EVALUATION of preliminary work
23 Collections of material Catalogues
24 Data file on collections of material
25 Determination of places where inquiries are to be carried out and of experts to be interrogated
26 Assignment of tasks to experts
Phase V

SLIDE 17 a

CAPABILITY

= LASTING PHYSICAL AND PSYCHICAL ABILITY OF A PERSON TO ACQUIRE OR APPLY KNOWLEDGE AND SKILLS

KNOWLEDGE

= ATTAINMENTS NECESSARY FOR CARRYING OUT A FUNCTION (PRIMARILY THEORETICAL)

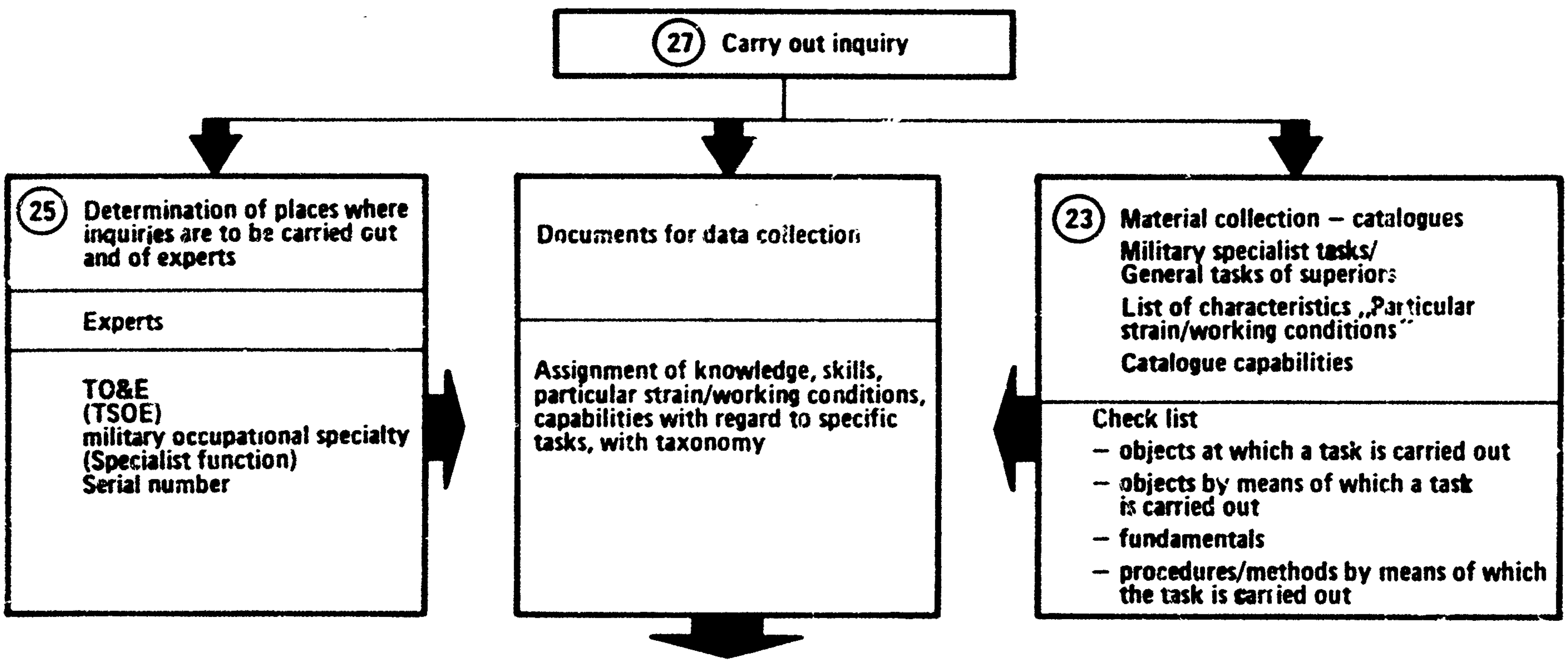
SKILLS

= ABILITIES NECESSARY FOR CARRYING OUT A FUNCTION (PRIMARILY PRACTICAL)

SLIDE 18

Analysis of Requirements Which Must be Fulfilled Phase V

Inquiry on knowledge and skills, assignment of characteristics „Particular strain/working conditions", assignment of capabilities

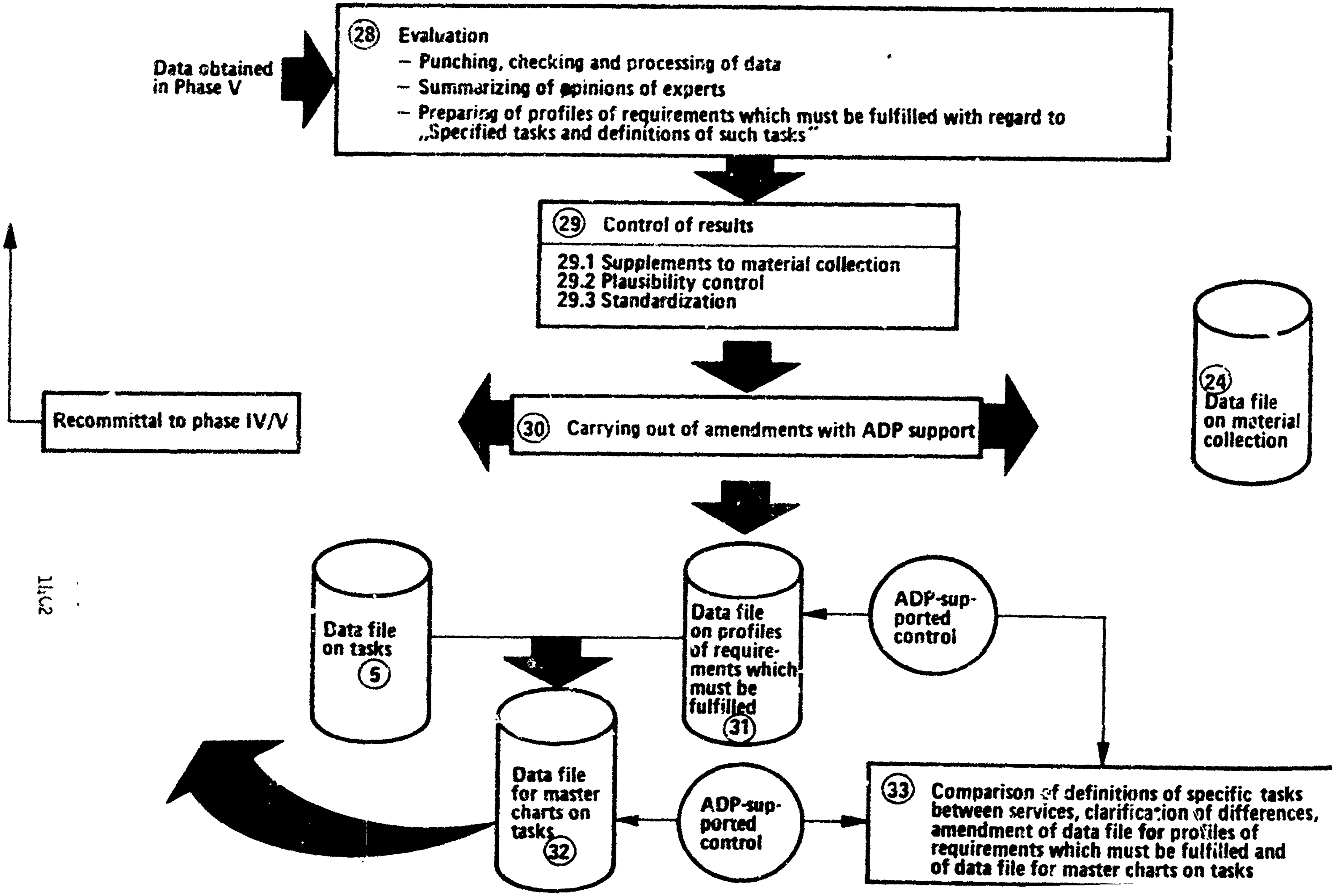


(Evaluation)
Phase VI

SLIDE 19

Analysis of Requirements Which Must Be Fulfilled Phase VI (Evaluation)

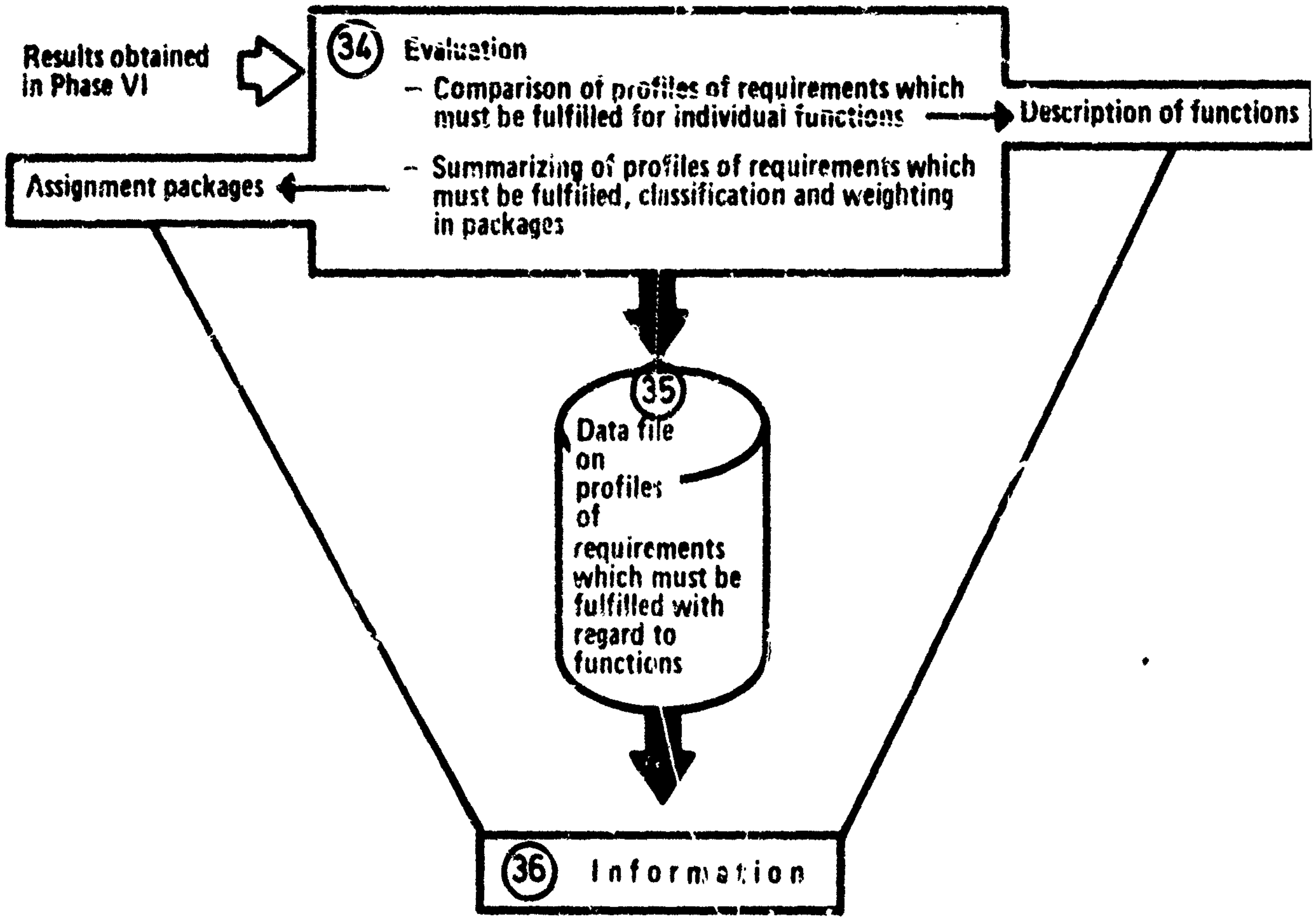
Preparing of profiles for requirements which must be fulfilled with regard to definitions of specific tasks



SLIDE 20

Funktionsanalyse Personalstruktur Phase VII (Analysis of functions pertaining to a job)

(Evaluation and determination of assignment packages)



SLIDE 21

<u>ALTOGETHER</u> NECESSARY:

- 104 CATALOGUES OF QUESTIONNAIRES ON TASKS, EACH CATALOGUE CONTAINING UP TO

  . 200 TASKS

BY THE <u>END OF 1976</u>

- 28 CATALOGUES AND LISTS CONTAINING DEFINITIONS OF SPECIFIC TASKS

  WERE COMPLETED

BY THE <u>END OF 1977</u>, PRESUMABLY

- TWO THIRDS OF ALL CATALOGUES WILL BE COMPLETED
- A DESCRIPTION OF 50 % OF ALL JOBS / FUNCTIONS WILL BE POSSIBLE

SLIDE 22

FOR INQUIRY ON ACTUALLY PERFORMED TASKS

ARE NECESSARY:

- 54,000 SOLDIERS

- 1,800 INTERROGATION DAYS (1 TEAM)

PLANNED FOR 1978:

- 5 TEAMS

1981 PRESUMABLY

- TERMINATION OF THIS INQUIRY

SLIDE 23

INTERMEDIATE RESULTS

POSSIBILITY OF CALLING AT ANY TIME INTER ALIA:

- TASK DESCRIPTIONS ACCORDING TO GENERIC TERMS
- JOB DESCRIPTIONS ACCORDING TO PRESENT ORGANIZATION
- MANY INDIVIDUAL QUESTIONS (COMPARE SLIDES 24 AND 25)

SLIDE 24

STANDARD OF FUNCTIONS

1. Identification, for example:

   number in TO&E, designation in TO&E
   appointment
   own forces / means
   position as superior
   security classifications, clearance
   knowledge of foreign languages (according to code of the Federal Armed Forces)

2. Descriptive characteristics, for example:

   main tasks
   operational command and control tasks
   tactical command and control tasks
   general tasks of superiors
   knowledge, skills
   knowledge of objects
   knowledge of fundamentals
   environmental / working conditions

SLIDE 25

3. <u>Employment, training</u>, for example:

necessary previous military employment
desired previous military employment
necessary courses
possible further military employment (horizontal)
possible further military employment (vertical)
transition to the civilian sector

4. <u>Evaluating characteristics</u>, for example:

minimum enlistment period of reenlisted men
area of responsibility
level of responsibility

OCCUPATIONAL TASK FACTORS FOR INSTRUCTIONAL SYSTEM DEVELOPMENT

By
William Stacy
Nancy Thompson
Sq Ldr David Thomson

Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Brooks AFB, TX

I. INTRODUCTION

The Air Force Human Resources Laboratory has been engaged in a long range research effort to determine task training requirements based upon occupational survey data. The methodology being developed is in support of the Air Force Instructional Systems Development (ISD) program. Specifically, occupational task factors are being developed and utilized to provide data in meeting the ISD requirements of analyzing the system or job requirements and defining educational training needs based upon job performance requirements.

The primary consideration for including or excluding a task for training has been based upon occupational survey data of the probability of airmen performing certain tasks in their first job assignment. Changes in curriculum design based upon these types of studies have saved the Air Force millions of dollars by eliminating training on tasks performed by low percentages of first-term airmen in the field. However, the probability of the performance of a task may not be the only factor to be considered in establishing training requirements. For example, if the consequences of inadequate performance of a task are hazardous or costly to human life or property, that task should be considered for training regardless of the frequency of performance. In recent years, AFHRL has hypothesized and researched a number of task training factors important in determining how much emphasis should be placed on tasks for training.

The basic theory behind the Air Force task training factor research was conceived and reported by Christal (1970). In the design of the research a number of current task factors have been identified which can provide training development personnel with an objective procedure for determining the task's priority for training. Data are currently being collected on the following task factors:

a. Percent members performing. The percent of airmen in the career specialty performing the task in their jobs.

b. Task difficulty. The time required to learn to perform the task satisfactorily.

c. Consequences of inadequate performance. The perceived consequences if the task is incorrectly performed, considered in terms of destroyed material, wasted time, injury, or loss of life.

d. Task delay tolerance. A measure of how much delay can be tolerated between the time the airman becomes aware the task is to be performed and the time he must commence doing it.

e. Field recommended training emphasis. A measure of the task's recommended formal training emphasis (either school or OJT), based upon the ratings of tasks by 7- and 9- skill-level field NCOs.

f. School training emphasis. A measure of the task's current training emphasis in residenc training, based upon the ratings of tasks by course instructors.

The task training factor methodology has many potential applications. For each of the task factors under consideration, the task from occupational surveys can be ordered in sequence. For example, the occupational survey tasks can be ordered in descending sequence based upon the arithmetic mean ratings of consequences of inadequate performance. Information may also be provided comparing school training emphasis with field recommended training emphasis. In addition, using multiple regression analysis, the task factors can be used as predictor variables to capture the judgments of school training personnel and recommended training emphasis from the field. The task factors can be applied in the selecting of tasks for training courses, the development of specialty training standards, the validation of current training courses, or for the redesigning of an existing course as a result of changes in course length or changes in the career specialty. The task training factor methodology does not provide all answers to training course decisions; the methodology is primarily as an advanced aid to course design and is subject to override by other training considerations as required.

The development of task training priority factors was described by Mead, (1975) at the 17th annual conference of the Military Testing Association. In Mead's paper, procedures for validating task factors were illustrated by two studies. In one validation study (Mial & Christal, 1974), one hundred-ninety first-term airman tasks in the Medical Services (902X0) career specialty were placed on 4" X 6" cards and were rank ordered by curriculum specialists according to their priority for resident technical training. The mean rank values served as the criterion measure and were predicted by the task factors using multiple regression analysis. This policy capturing procedure resulted in a four variable task factor equation which correlated $R = .88$ with those of the curriculum specialists.

A second validation study was conducted by Mead, (1975) using the Law Enforcement (811XX) specialty. Training curriculum specialists rank ordered a representative sample of 165 apprentice and journeyman level tasks as to their priority for formal training. The mean rank values for each task were used as the criterion values. The training priority policy of the curriculum specialists was predicted using four task factors and derivations of these factors. The equation provided training priority values which correlated $R = .95$ ($R^2 = .91$) with the criterion values. These studies strongly indicated that the task training policies of curriculum specialists could be duplicated mathematically using the task factor prediction model.

II. CURRENT TASK FACTOR RESEARCH PROCEDURES

Two types of scales are currently applied in the collection of ratings on task difficulty, task delay tolerance, and consequences of inadequate performance. Relative task factor scales have been used to rate each task in a career specialty relative to the other tasks in that specialty. Benchmark or task-anchored scales have also been designed which are used to rate the tasks in one specialty compared with prescribed levels of tasks in other Air Force specialties.[1]

A major limitation of the relative scales is that the tasks in a career specialty are only rated in a context with the other tasks in that specialty.

> For example, using a relative scale, a cook may rate "salting the meat" very high in consequences of inadequate performance. Yet, in a task-anchored scale, the consequences of inadequate performance of "salting the meat" may not seem very serious when compared with inadequate performance of other Air Force tasks.

In the past several years AFHRL has demonstrated considerable success in the development of task-anchored rating scales. The purpose of the task-anchored scales is to permit comparisons of tasks within one career specialty with representative tasks performed in other Air Force specialties. Task-anchored scales have been developed for three task factors: task difficulty, consequences of inadequate performance, and task delay tolerance. The scales were also developed to represent three aptitude areas: a combined administrative or general requirement; electronic aptitude; and mechanical aptitude. An illustrate of an administrative/general aptitude task anchored scale is presented in Appendix A. The scale is composed of 27 tasks, subdivided into nine subgroups of

---

[1] The term benchmark scales has recently been changed to task-anchored scales to provide additional clarity.

three tasks. Each subgroup represents one of nine levels on the scale. In actual use, raters compare each of the tasks in their specialty against the 27 tasks in the benchmark scale. For each task in his specialty the rater will decide which level and subgroup of three tasks are the most similar on the factor being considered. The development of the benchmark scales for the Administrative, General aptitude area has been reported by Goody (1976). The development of the electronic and mechanical benchmark scales has also been completed but not reported.

In Appendix B, inter-rater agreement of relative and task-anchored (benchmark) scale ratings is reported using the intra-class correlation technique (Lindquist, 1953, p. 361). From these reliability estimates of mean ratings it appears that the task factors used in the studies are generally stable and reliable measures. It has been demonstrated for the most part that the reliability estimates of mean ratings obtained from experienced NCOs are acceptably high and consistent.

In Appendix C, for an additional comparison on rater agreement between relative and task-anchored (benchmark) scales, four career specialties have been surveyed with both scales and their sample size adjusted to a common N = 50. It appears generally that the benchmark scale raters were as good or slightly better in their agreement with each other than the relative scale raters. The percentage of deviant raters who were deleted from the study because they used the scale upside down or did not properly follow instructions, was noticeably less for the benchmark scale raters.

For several reasons, future research studies will probably use benchmark scales in collecting task factor data. The benchmark scales are more advantageous than relative scales because they provide more information for making training decisions between career specialties, the benchmark scales appear to provide as good or better agreement among raters than relative scales, and require fewer case deletions. The benchmark scales can also be used to rank order training priorities of tasks within a single career specialty.

Appendix D, provides correlations of school and field training emphasis versus the task factors for twelve career specialties. The task factor data were collected with relative scales. One important correlation in Appendix D is the zero-order correlation between school training emphasis and field recommended training emphasis. In some of the career specialties there was fairly high correlation between what the field recommends to be taught on job tasks and what the school is currently teaching. In other career specialties the correlations between school and field training emphasis were not as high. In interpreting Appendix D, it appears that when correlations between school and field training emphasis were low, the correlation of school emphasis with task factors were also relatively low compared to correlations of field emphasis with task factors. In such a case, it would seem that

three tasks. Each subgroup represents one of nine levels on the scale. In actual use, raters compare each of the tasks in their specialty against the 27 tasks in the task anchored scale. For each task in his specialty the rater will decide which level and subgroup of three tasks are the most similar on the factor being considered. The development of the task anchored scales for the Administrative, General aptitude area has been reported by Goody (1976). The development of the electronic and mechanical task anchored scales has also been completed but not reported.

In Appendix B, inter-rater agreement of relative and task-anchored (benchmark) scale ratings is reported using the intra-class correlation technique (Lindquist, 1953, p. 361). From these reliability estimates of mean ratings it appears that the task factors used in the studies are generally stable and reliable measures. It has been demonstrated for the most part that the reliability estimates of mean ratings obtained from experienced NCOs are acceptably high and consistent.

In Appendix C, for an additional comparison on rater agreement between relative and task-anchored (benchmark) scales, four career specialties have been surveyed with both scales and their sample size adjusted to a common N = 50. It appears generally that the bechchmark scale raters were as good or slightly better in their agreement with each other than the relative scale raters. The percentage of deviant raters who were deleted from the study because they used the scale upside down or did not properly follow instructions, was noticeably less for the task anchored scale raters.

For several reasons, future research studies will probably use task-anchored scales in collecting task factor data. The benchmark scales are more advantageous than relative scales because they provide more information for making training decisions between career specialties, the task anchored scales appear to provide as good or better agreement among raters than relative scales, and require fewer case deletions. The task anchored scales can also be used to rank order training priorities of tasks within a single career specialty.

Appendix D, provides correlations of school and field training emphasis versus the task factors for twelve career specialties. The task factor data were collected with relative scales. One important correlation in Appendix D is the zero-order correlation between school training emphasis and field recommended training emphasis. In some of the career specialties there was fairly high correlation between what the field recommends to be taught on job tasks and what the school is currently teaching. In other career specialties the correlations between school and field training emphasis were not as high. In interpreting Appendix D, it appears that when correlations between school and field training emphasis were low, the correlation of school emphasis with task factors were also relatively low compared to correlations of field emphasis with task factors. In such a case, it would seem that

the school was giving less consideration to the task factors than was the field. However, low correlations between school and field ratings do not necessarily mean inadequacies in school curriculum. The schools are operating under a number of constraints as to what can be taught in resident training and field ratings are influenced by consideration of OJT as well as resident training.

The analysis and display of Air Force task factor data is best accomplished through the CODAP analysis system. CODAP is a comprehensive set of computer programs for analyzing and reporting occupational information collected with job inventories (Christal & Weissmuller, 1976). One of the most frequently used CODAP programs to illustrate task factor data is called FACSUM. Appendix E presents a FACSUM difference description between school and field training emphasis in the Medical Services (902X0) career specialty. The top of the description shows tasks which have the largest difference between what is recommended by the field for training and what is currently being taught in the school. Tasks are listed in descending sequence of these differences. The tasks listed at the bottom of the description are those on which the school emphasis was greater than the emphasis recommended by the field.

The task factor data developed in the Medical Services study proved to be extremely valuable in aiding ISD and training development personnel in making changes in the resident course based upon the task factor information.

The power of the task factor data in aiding training development personnel can be further evidenced in the Dental Specialists (981X0) training course. From examining a FACSUM difference description between school and field recommended training emphasis there were two tasks which appeared at the top of the description.

TASK	SCHOOL	FIELD	DIF FER	CON SEQ	TSK DELAY	TSK DIFF	% PERF
H273 Fabricate Custom Impression Trays	5.7	1.3	4.4	3.2	5.9	3.7	7.7
G240 Pour and trim dental casts	5.7	3.2	2.5	3.1	4.8	3.1	23.6

Based upon the difference in training emphasis between school and field and the low percentages of first-term airmen performing the tasks the school investigated the status of these two tasks. It was determined that the two tasks were being performed primarily in a laboratory function and were no longer considered as routine tasks in the field. The training of these two tasks in the resident course was suspended representing a savings of approximately fourteen and one-half hours in the training cousse plan of instruction (POI).

III. SUMMARY

The task training factor methodology developed by AFHRL has demonstrated many applications in determining the training priorities of tasks based upon occupational survey data. The primary benefits of this research have been in the use of task factor data for designing training courses, the development of specialty training standards, the validation of current training courses, and for the redesigning of an existing course as a result of changes in course length or changes of the career specialty. The development of the task-anchored scales has opened an additional area of research for exploring differences in training emphasis and priority between Air Force career specialties. The recent collection of recommended task training requirements from the field and of information concerning current training emphasis in the schools has provided meaningful comparisons which can be used in determining the priorities of tasks for training. In attempting to capture the policies of field and training school personnel in establishing training requirements, task factors have been used as predictor variables in multiple regression analysis. The task factor methodology although still in the developmental process, has been demonstrated to be extremely valuable in the curriculum design.

REFERENCES

Christal, R. E. Implications of Air Force occupation research for curriculum design. In B. B. Smith & J. Moss, Jr (Eds.), Report of a seminar: Process and techniques of vocational curriculum development: Minnesota Research Coordinating Unit for Vocational Education, University of Minnesota, Minneaplis, MN, April 1970, 27-61.

Christal, R. E., & Weissmuller, J. J. New CODAP programs for analyzing task factor information. AFHRL-TR-76-3. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, May 1976 AD-A026 121.

Goody, K. Task factor benchmark scales for training priority analysis. Overview and developmental phase for administrative/general aptitude area. AFHRL-TR-76-15. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1976. AD A025 847.

Goody, K., & Watson, W. J. Task factor benchmark scales for use in determining task training priorities. Paper presented at the 17th Annual Conference of the Military Testing Association, U. S. Army, Indianapolis, IN, 16-19 September 1975.

Lindguist, E. F. Design and Analysis of Experiments in Psychology and Eduction, 1953.

Mead, D. F. Determining training priorities for job tasks. Paper presented at the 17th Annual Conference of the Military Testing Association, U. S. Army, Indianapolis, IN, 16-19 September 1975.

Mial, R. P., & Christal, R. E. The determination of teaining priority for vocational tasks. Proceedings, Psychology in the Air Force Symposium. USAF Academy, April 1974, 29-33.

*Presented by the Military Association Conference, 17-20 October 1977.

APPENDIX A

TASK DELAY TOLERANCE

(Administrative/General)

DEFINITION

The Task Delay Tolerance of a task is a measure of how much delay can be tolerated between the time the airman becomes aware the task is to be performed and the time he must commence doing it.

BENCHMARK SCALE

Level 1 – Least Tolerance of Delay – Must do immediately

Use artificial respiration to restore breathing of accident or fire victims (Fire Protection Specialist)
Issue scramble orders to fighter aircraft (Command and Control Specialist)
Assist during treatment of cardio-respiratory failure in operating room (Operating Room Specialist)

Level 2

Quell disturbances involving military personnel (Security Specialist)
Identify tablets, capsules or liquids involved in poisoning cases (Pharmacy Specialist)
Operate safety console at missile control center during hazardous operations (Missile Safety Specialist)

Level 3

Inspect runway for foreign objects (Air Operations Specialist)
Administer anaesthesia in dental surgery (Dental Specialist)
Adjust airborne radio receivers to obtain readable signals (Radio Operator)

Level 4

Question suspects or witnesses (Security Specialist)
Perform colony counts on bacteria to estimate type and level of infection (Medical Laboratory Specialist)
Maintain proper temperature of food storage areas (Cook)

Level 5

Identify military vehicles, installations or activities in visual photographs (Intelligence Operations Specialist)
Proofread or correct teletype tape or page copies (Communications Center Specialist)
Prepare daily weather maps (Weather Forecaster Specialist)

Level 6

Operate computer remote inquiry terminals (Computer Operator)
Purge or clear chemical lines in film developing machines (Still Photographic Laboratory Specialist)
Service and maintain dental high-speed drilling equipment (Dental Laboratory Specialist)

Level 7

Monitor workload reporting systems (Manpower Specialist)
Brief personnel on state or local motor traffic laws (Safety Specialist)
Draw up work rosters for taxi operators or drivers on large Air Force base (Programs and Work Control Specialist)

Level 8

Write item identification descriptions and specifications for catalogues (Procurement Specialist)
Interview or hire civilian personnel (Supply Service Specialist)
Prepare and analyse work flow process charts (Management Engineering Specialist)

Level 9 – Most Tolerant of Delay – Do when ready

Review or select books or publications for unit library (Administration Specialist)
Research and write feature stories in Air Force publications (Information Specialist)
Clean teeth of animals (Verterinary Specialist)

APPENDIX B

RELIABILITY OF TASK FACTOR RATINGS

Air Force Specialty	Scale	Field Recommended Emphasis R_{kk}	Field Recommended Emphasis N Rater	Field Recommended Emphasis % Deleted	Current School Emphasis R_{kk}	Current School Emphasis N Rater	Current School Emphasis % Deleted	Consequences of Inadequate Performance R_{kk}	Consequences of Inadequate Performance N Rater	Consequences of Inadequate Performance % Deleted	Delay Tolerance R_{kk}	Delay Tolerance N Rater	Delay Tolerance % Deleted	Task Difficulty R_{kk}	Task Difficulty N Rater	Task Difficulty % Deleted
293X3 Radio Operator	Relative	.98	199	16	.99	9	0	.90	36	20	.91	38	26	.87	59	26
	Benchmark	-	-	-	-	-	-	.91	47	8	.94	42	16	.91	41	16
304X0 Radio Relay	Relative	.99	194	7	.92	11	0	.77	36	3	.90	60	20	.93	89	11
	Benchmark	-	-	-	-	-	-	.92	37	5	.89	39	0	.94	63	7
304X4 Ground Radio Equip.	Relative	.99	315	5	.99	6	0	.88	36	7	.76	49	16	.96	108	11
	Benchmark	-	-	-	-	-	-	.90	63	5	.89	49	16	.96	55	9
651X0 Procurement	Relative	.97	295	8	-	-	-	.87	58	5	.90	51	19	.95	77	12
	Benchmark	-	-	-	-	-	-	.86	59	12	.93	59	17	.96	54	[illegible]
531X5 Non Destructive Inspection	Benchmark	.99	178	8	-	-	-	.91	57	7	.92	50	25	.96	62	[illegible]
328X3 Electronic Warfare	Relative	.99	248	19	.94	9	0	.80	45	2	.74	43	9	.90	39	9
472X2 Vehicle Mechanic	Relative	.99	243	16	.96	11	0	.77	28	15	.47	24	29	.96	84	16
552X5 Plumbing	Relative	.98	125	6	.98	6	0	.92	77	6	.83	45	27	.96	75	15
672X1 Gen Accounting	Relative	.97	76	11	.97	11	0	.83	169	9	.86	108	16	.91	75	13
672X2 Disbursement Acct	Relative	.99	135	9	.98	18	0	.83	109	9	.86	108	16	.91	75	13
902X0 Medical Services	Relative	.99	277	27	.97	20	0	.94	93	7	.95	95	4	.96	58	-
906X0 Medical Administrative	Relative	.99	270	10	.89	7	0	.92	103	2	.94	95	9	.96	73	6
913X0 Physiological Train.	Relative	.98	68	16	.96	11	0	.87	30	0	.77	34	30	.96	71	-
981X0 Dental Specialist	Relative	.99	85	6	.97	20	0	.96	63	3	.95	44	0		-	20

APPENDIX C

INTER-RATER RELIABILITY BASED ON A SAMPLE OF 50 (R_{5050})

Air Force Speciality	Consequences of Inadequate Performance		Task Delay Tolerance		Task Difficulty	
	Relative Scale	Benchmark Scale	Relative Scale	Benchmark Scale	Relative Scale	Benchmark Scale
293X3 Radio Operator	.93	.94	.94	.96	.90	.94
651X0 Procurement Specialist	.85	.85	.90	.92	.94	.95
304X0 Radio Relay Equipment Repairman	.83	.94	.93	.91	.92	.95
304X4 Ground Radio Communication Equipment Repairmen	.87	.88	.79	.90	.95	.95

APPENDIX D

CORRELATIONS BETWEEN TASK FACTORS AND TRAINING EMPHASIS

Air Force Specialty	Training Emphasis	Percent Members Performing (1-24 Months)	Consequences of Inadequate Performance	Delay Tolerance	Task Difficulty	R^2 Full Model*	Correlation School vs Field
293X3 Radio Operator	School	.46	.22	-.53	-.04	.55	
	Field	.82	.38	-.82	-.00	.90	.62
304X0 Radio Relay	School	.65	.19	-.46	.05	.52	
	Field	.73	.44	-.80	-.11	.86	.66
304X4 Ground Radio	School	.60	.09	-.22	-.02	.50	
	Field	.77	.20	-.35	.03	.73	.63
328X3 Electronic Warfare	School	.39	.21	-.16	-.05	.29	
	Field	.79	.39	-.52	-.49	.81	.48
472X2 Vehicle Mechanic	School	.87	.17	-.12	-.32	.85	
	Field	.87	.15	-.12	-.30	.91	.93
552X5 Plumbing	School	.70	.00	-.11	-.45	.67	
	Field	.79	.31	-.32	-.35	.89	.75
672X1 General Accounting	School	.64	.17	-.20	-.16	.54	
	Field	.70	.24	-.20	-.16	.72	.76
672X2 Disbursement Accounting	School	.69	.06	-.48	-.27	.53	
	Field	.85	.00	-.62	-.43	.87	.76
902X0 Medical Services	School	.75	.80	-.22	-.39	.62	
	Field	.35	.70	-.73	.20	.72	.29
906X0 Medical Administrative	School	.63	.13	-.33	-.25	.56	
	Field	.48	.37	-.39	-.04	.59	.71
911X0 Physiological Training	School	.64	.16	-.53	-.09	.63	
	Field	.74	.32	-.76	-.26	.84	.76
981X0 Dental Specialist	School	.89	.14	-.60	-.58	.81	
	Field	.85	.38	-.79	-.61	.89	.85

NOTE - *Predictors = Percent performing, Consequences, Delay, Difficulty, Percent Time spent, Average Percent Grade, and these predictors squared.

APPENDIX E

FACSUM Difference Description Between Field and School Training Emphasis Medical Services 902X0

Task		Field Recommended Emphasis	Current School Emphasis	Difference (Field minus School)	Percent Performing (1-24 months)	Consequences	Delay Tolerance	Task Difficulty
I 419	Suture Lacerations	6.31	.00	6.31	23.6	5.83	3.52	6.55
G 295	Insert Oral Airways	6.61	.30	6.31	24.3	7.10	2.06	5.36
I 378	Administer Primary Care at Scene of Accidents	6.99	1.35	5.64	23.9	7.41	1.87	6.44
I 391	Drive Ambulances or Ambuses	5.62	.00	5.62	20.8	5.95	4.32	4.96
I 411	Remove Sutures	5.56	.00	5.56	33.0	4.40	5.96	4.39
F 179	Prepare Plaster of Paris, Cotton, or Other Materials for Fabrication of Casts	5.70	.20	5.50	23.5	4.19	5.00	4.92
F 174	Prepare Patients for Minor Surgery	6.08	.60	5.48	40.5	4.99	4.74	4.80
G 252	Apply Pneumatic Splints	5.48	.05	5.43	11.4	5.18	4.34	4.45
G 254	Apply Short Leg Plaster Casts	5.39	.00	5.39	15.5	5.21	5.39	5.42
G 253	Apply Short Arm Plaster Casts	5.39	.00	5.39	16.8	5.18	5.40	5.40
G 280	Draw Blood from Patients	5.38	.00	5.38	29.5	5.58	4.46	5.60
G 248	Apply Make-Shift Splints	5.59	.25	5.34	18.9	5.25	4.21	4.84
	Tasks in Middle of Description Omitted							
H 366	Orient New Patients	4.42	4.90	-.48	51.7	3.53	5.23	3.78
G 243	Apply Heat by Hot Water Bottles	4.90	5.40	-.50	32.8	4.85	5.37	3.39
G 296	Instruct Patients in Crutch Walking	4.95	5.50	-.55	32.8	4.82	5.42	4.60
G 312	Perform Oral Hygiene	5.21	5.85	-.64	45.4	4.06	5.16	4.18
G 202	Administer Complete Bed Baths	5.55	6.25	-.70	51.2	3.00	5.40	3.96
H 361	Make Occupied Beds	5.07	6.00	-.93	54.6	3.57	5.47	4.02
G 309	Participate in Team Conferences	4.10	5.10	-1.00	37.5	3.71	6.04	4.46
G 284	Give Back Rubs	4.54	5.55	-1.01	45.3	3.14	6.19	3.78
H 362	Make Postoperative or Recovery Beds	4.92	6.00	-1.08	51.0	3.13	4.91	3.65
G 199	Administer Bed Pans or Urinals	4.74	6.15	-1.41	66.4	2.65	4.19	2.86
H 363	Make Unoccupied Beds	4.56	6.10	-1.54	60.7	2.31	6.29	3.11
H 373	Serve Meal Trays	3.63	5.40	-1.77	53.9	3.44	5.07	3.14

REPORT OF STEERING COMMITTEE

&

GENERAL BUSINESS MEETING (1977)

1. HARRY H. GREER AWARD:

Steering Committee approved presentation of the swards to Dr. William Moonan, Naval Personnel and Research Development Command, and to John A. Burt, U. S. Coast Guard Institute. (Texts are atch).

2. ARTICLE III OF THE BY-LAWS

The Steering Committee voted not to change the wording of Article III. This article deals with membership.

3. ARTICLE VII OF THE BY-LAWS

The Steering Committee and the General Membership approved a change of Section B of the By-Laws from the present wording of:

B. The annual Conference of the association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment in order of the following rotating schedule:

United States Army
United States Marine Corps
United States Navy
United States Air Force
United States Coast Guard;

Henceforth to read:

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The coordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

4. COORDINATING AGENCIES 1978-1982

In conformance with Article VII, Section B the hosting sites for the above periods are:

1978	Oklahoma City - Coast Guard Institute
1979	San Diego - Navy Personnel Research and Development Center
1980	Toronto, Canada - Canadian Forces Personnel. Applied Research Unit
1981	Ft Eustis, Virginia - U. S. Army
1982	Pennsacola, Florida - U. S. Navy, Program Development Center

MILITARY TESTING ASSOCIATION

HARRY F. GREER AWARD

TO

JOHN A. BURT

To you, John A. Burt, the Military Testing Association owes its current level of prominence and excellence. Your personal efforts over the last ten years have helped insure the continuance of the MTA as a functioning and growing organization, and have helped produce interesting, informative, and important conferences over the years. Your continuing involvement in the activities of the steering committee and the overall management of the MTA have been appreciated. We all realize that much of this service to the MTA has been accomplished at the cost of difficult and time-consuming personal effort.

This award for outstanding service is most appropriately given in the name of Harry F. Greer, as you have exemplified his aims and have carried out his intentions in forming the MTA. This award is made with the gratitude, friendship, and regard of all associated with the Military Testing Association.

MILITARY TESTING ASSOCIATION

HARRY F. GREER AWARD

TO

DR. WILLIAM MOONAN

The Harry Greer Award is hereby presented to Dr. William Moonan of the Naval Personnel Research and Development Center for your consistent and lasting contributions to the purposes of the Military Testing Association.

Year after year, your dedication to the scientific principles which underlie the assessment of individuals has inspired each of us while the integrity and innovation of your work has set an example for all. Your original contributions, in the statistical treatment of assessment data are numerous.

The prodigous volume and inventive character of your work is of credit to yourself, to the Navy and to the entire community of military personnel assessment. For this we thank you. Therefore this award is made with gratitude, friendship and regard of all associated with the Military Testing Association.

BY-LAWS OF THE MILITARY TESTING ASSOCIATION*

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

* As approved at the 1977 General Meeting of The Association, 21 Oct 77, San Antonio, Texas

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. A majority of the members of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term ot the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to insure notices for conferences. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

* B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The coordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

C. The annual conference of the Association shall be held at a time and place determined by the coordinating agency. The membership of the association shall be informed at the annual conterence of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authoriztion of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 2 November 1973).

ATTENDEES

ADAMS, William, Jr.
Chief of Naval Educ & Training
Naval Air Station
Pensacola, FL 32508

ALBECK, Robert D
Naval Tech Trng Center
Corry Station
Pensacola, FL 32511

ALBERT, Walter G
AFHRL/SM
Brooks AFB, TX 78235

ALLEY, William E
AFHRL/ORE
Brooks AFB, TX 78235

AMMERMAN, Harry L, Dr
Center for Vocational Educ
1960 Kenny Road
Columbus, Ohio 43210

ANDERSON, H. R.
USCG AVTRACEN
Bates Field,
Mobile, AL 36608

ARCHER, Wayne B
AFHRL/ORA
Brooks AFB, TX 78235

ARIMA, James K
3115 Martin Rd
Carmel CA 93921

ASA-DORIAN, Paul V, Sr
Fleet Anti-Submarine Wareface Tng Cen
San Diego, CA 92147

AVERSANO, Frank M, Dr
US Army Trng Support Center
Ft Eustis, VA 23604
ATTN: ATTSC-TI-TD

AVERY, Doris E
Army Educ Center
P.O. Box 3533
Ft Wainwright, Alaska

BACHTEL, Mary A
Naval Educ & Trng Progr Deve Cntr
Ellyson PD-9
Pensacola, FL 32509

BAISDEN, Annette
Psychology Dept
Naval Aerospace Med Research Lab
Pensacola, FL 32508

BAKER, J. G. Capt
Chief of Naval Educ & Trng/N-8
NAS, Pensacola, FL 32508

BARAN, Harry A.
AFHRL/ASR
Wright-Patterson AFB, OH 45433

BAREFIELD, Ruth B
USAAVNC
Ft Rucker, AL 36362

BARLOW, Bruce H, Capt
Directorate of Tng Dev
US Army Inf School
Ft Benning, GA 31905

BAYMOR, Michael
Directorate of Evaluation
USAARMS
Ft Knox, KY 40121

BECK, Vera J
Education Center - DPCA
Ft Ord, CA

BEGLAND, Bob Capt
3033 Stillwood Ct
Tallahassee, FL 32303

BELISLE, Michael
3507 ACS
119 Robinhood
San Antonio, TX 78209

BELL, M. Herman, LTC
Directorate of Tng Dev
US Army Inf School
Ft Benning, GA 31905

BELLANTONI, Maria, Dr
US Army Trng Dev Inst
Bldg 1514
Ft Eustis, VA 23604

BERCER, B. Michael
USA MIL PER CEN
2461 Eisenhower Ave
Alexandria, VA 22331

BERGER, Daniel L, Sr A
AFHRL/SM
Brooks AFB, TX 78235

BERGMANN, Joe A
AFHRL/ORA
Brooks AFB, TX 78235

BIALEK, Hilton M, Dr
Human Resources Rsch Organ (HUMRRO)
27857 Berwick Dr
Carmel, CA 93923

BIEDIGER, Rosalie A
AFHRL/SM
Brooks AFB, TX 78235

BIRDALL, Walter A
Naval Educ & Trng Program Dev Cen
Ellyson, PD
Pensacola, FL 32509

BLACK, Doris E
AFHRL/SM
Brooks AFB, TX 78235

BLAND, Raymond D, CMDR
USCG Training Center
Cape May, NJ

BOLDOVICI, John A, Dr
HUMRRO
PO Box 293
Ft Knox, KY 40121

BONDARUK, John
Nat'l Security Agency
Ft George G Meade, MD 20755

BORTNER, Don E, SrA
AFHRL/NA
Brooks AFB, TX 78235

BOTTENBERG, Robert A, Dr
AFHRL/SM
Brooks AFB, TX 78235

BOUZA, Duane J
Base Educ Cen (DPT)
Offutt AFB, NE 68113

BOYD, Joseph L, Jr
Educational Testing Service
10 Pine Knoll Dr
Lawrenceville, NJ 08648

BOYD, Richard D
MILPERCEN USA
Alexandria, VA 22314

BRAND, John S
Directorate of Evaluation
US Army ADMINCEN
Ft Benjamin Harrison, Ind 46249

BROKAW, Leland D, Dr
AFHRL/PE
Brooks AFB, TX 78235

BROWN, Fred W, MSgt
AFHRL/SM
Brooks AFB, TX 78235

BREWER, David B, Ensign
Commandant (G-RT/81)
USCG HQ
Washington, DC 20590

BRITTAIN, Clay V
Army Trng Support Center
Ft Eustis, VA 23185

BUCK, C. W.
U.S. Cost Guard Inst/Dev Div
P.O. Substation 18
Oklahoma City, OK 73169

BURT, John A
USCG Institute
P.O. Substation 18
Oklahoma City, OK 73169

BOYD, Richard D
MILPERCEN USA
Alexandria, VA 22314

BURTCH, Lloyd D
AFHRL/ORA
Brooks AFB, TX 78235

CALL, W. B.
Education Center
Ft Ord, CA

CALLAHAN, M D
NCTAP Dept NODAC
Washington, D.C. 20370

CAMPBELL, Charlotte H
HUMRRO
P.O. Box 293
Ft Knox, KY 40121

CAMPBELL, Roy C
HUMRRO
P.O. Box 293
Ft Knox, KY 40121

CAPATOSTO, Richard P, L/C
Office of Manpower Utilization
MCDEC (OMU)
Hq Marine Corps
Quantico, VA 22134

CARPENTER, Gerald, L/C
CO CdN Forces Pers Applied Rsch Unit
1107 Avenue Road
Toronto, Ont, Canada M5n 2E4

CHIPPENDALE, Joan T
USCG Training Center
Govenors Island, NY 10004

CHRISTAL, Raymond E., Dr
AFHRL/OR
Brooks AFB, Texas 78235

CLARK, Herbert J, Dr
AFHRL/XR
Brooks AFB, Texas 78235

CLOW, Gerald R
OMY
Lackland AFB, TX 78236

COLEMAN, Don
Training Dept, NAS Whiting Field
Milton, FL 32570

COOK, Paul, Capt
OMY
Lackland AFB, TX 78236

COOPER, Ronald S. SFC
US Amry Air Defense School
ATSA-TD-IT
Ft Bliss, TX 79916

CORMACK, Bruce, Cdr
Bruce Cormack Assoc, Lmtd
24 Glenridge Road
Ottawa, Ontario, Canada K2G2Z5

CORNELIUS, Edwin T, III, Dr
Assistant Professor
Ohio State University
404C W. 17th
Columbus, OH 43210

CORY, Charles H, Dr
Navy Personnel Rsch & Dev Cen
San Diego, CA 95152

COWAN, Douglas K
AFHRL/ORA
Brooks AFB, TX 78235

CRACKEL, Yvonne
Army Education Center
Ft Wainwright, Alaska 99701

CRISLER, Art M
ITED
Ft Eustis, VA 23604

CROW, Graham P, L/C
AFHRL/XR
Brooks AFB, TX 78235

CURRIER, Walter B, Capt
AFHRL/SM
Brooks AFB, TX 78235

CZUCHRY, Andrew, Dr
Dynamics Rsch Corp
60 Concord St
Wilmington, MA 01887

DAVIS, Chas
Office of Naval Rsch
Chicago, Ill
368 W. Alma Terrace
Gary, Illinois

DEVEAU, Larry
Chief of Naval Tech Trng Command
NAS Memphis
Millington, Tennessee 38054

DICKINSON, Richard W
Occupational Rsch Program
Industrial Engineering Dept
Texas A&M Univ
College Station, TX 77843

DITTMAR, Martin, Maj
OMY
Lackland AFB, TX 78236

DITULLIO, Paul N
OMY
Lackland AFB, TX 78236

DOLL, Richard E, Dr
Naval Aerospace Medical Rsch Lab
Pensacola, FL 32509

DOORLEY, Richard D
US Army MILPERCEN
200 Stovall St
Alexandria, VA 22332

DOW, Andrew N
USNETPDC
Bldg 923
Pensacola, FL 32509

DOWNEY, Hal W, LTC
Office of the Adjutant General
Washington, D. C. 20314

DREW, Richard
Central Test Site/PTEP
Dam Neck, VA 23452

DYER, Frederick N, Dr
Army Rsch Institute
P.O. Box 2089
Ft Benning, GA 31905

DUFFY Paul C
Marine Corps Inst
P.O. Box 1775
Washington, D.C. 20013

DRISKILL, Walter E, Dr
OMY
Lackland AFB, TX 78236

EARLES, James A
AFHRL/PE
Brooks AFB, TX 78235

EDWARDS, John O, Capt
AFHRL/ORE
Brooks AFB, TX 78235

ELLIS, John
NPRDC
San Diego, CA 92152

FARSON, Robert E., Jr., SrA
AFHRL/SM
Brooks AFB, TX 78235

FAST, Jonathan, Capt
AFHRL/ORS
Brooks AFB, TX 78235

FERGUSON, James E., Capt
US Coast Guard Inst.
P.O. Substation 18
Oklahoma City, OK 73169

FIEBIG, Gunter Flot Admiral
Amtschef Streitkrafteamz RADM
5300 Bonn-Bad Godesberg
Deutschherrenstrabe

FINSTUEN, Kenneth
AFHRL/ORE
Brooks AFB, TX 78235

FISCHL, M A, Dr
US Army Rsch Inst.
5001 Eisenhower Ave
Alexandria, VA 22333

FOLEY, Paul
Navy Personnel Rsch & Dev Cen
San Diego, CA 92152

EDSEL, Ford
US Army
172nd Infantry
Ft Greely, AK

FRESNE, Calvin C
AFHRL/SM
Brooks AFB, TX 78235

FULGHAM, Dan D, Col
AFHRL/CC
Brooks AFB, TX 78235

GALLAGHER, J. T, Capt
USN
Naval Aerospace Medical Research
Laboratory
Pensacola, Florida 32508

GALLOWA, Jacob
NETPDC
Corry Sta Code PD-2
Pensacola, FL 32511

GEORGE, Clay E, Prof
Dept of Psychology
Texas Tech Univ
Lubbock, Texas 79409

GILBERT, Arthur C F, Dr
US Army Rsch Inst
5001 Eisenhower Ave
Alexandria, VA 22333

GIORGIA, M. Joyce
AFHRL/ORS
Brooks AFB, TX 78235

GOLDBERG, Roger
DANTES, Ellyson Center
Pensacola, FL 32509

GOLDMAN, Lawrence A.
US Army MILPERCEN
Alexandria, VA 22331

GOODGAME, Doug, Dr
Industrial Engineering Dept
Texas A&M University
College Station, TX 77801

GOODWIN, James G
US Army Missile Mat Readiness Cmd
Restone Arsenal, AL 35809

GORMAN, Charles D, Capt
OMY
Lackland AFB, TX 78236

GORMAN, Steven
HQ Marine Corps
Code MPI-28
Wash DC 20380

GOTT, C. Deene
AFHRL/SM
Brooks AFB, TX 78235

GOULD, Richard B
AFHRL/PE
Brooks AFB, TX 78235

GREENWAY, Charles A
AFHRL/SM
Brooks AFB, TX 78235

GREGOIRE, Harv
SY 702 SETD
NATC
Patuxent River, MD 20670

GROSS, Jon L, Major
HQ USAF (RDXX)
Washington, D.C. 20330

GROVER, Martha S
Defense Intelligence School
Washington, D. C. 20374

GUINN, Nancy, Dr
AFHRL/PE
Brooks AFB, TX 78235

GUY, Robert L.
NAVEDTRAPRODEVCEN
CODE PD-3
Pensacola, FL 32509

HALL, Eugene R
Training Analysis & Eval Group
Dept of the Navy
Orlando, FL 32813

HALTRECHT, Ed, Dr
Ontario, Hydro HZ-D17
700 University Ave
Toronto, Ontario, Canada

HANNAMAN, David L
Kinton, Inc
4660 Kenmore Ave
Alexandria, VA 22304

HARMAN, Samuel A
Tng Div USA Facilities Engr
Support Agency
Ft Belvoir, VA 22060

HARRIS, James H
HUMRRO
P.O. Box 293
Ft Knox, KY 40121

HARRIS, Marilyn E, Dr
806 Metropolitan Bldg
Flint, MI 48502

HARRIS, Ward, Dr
US Army Training Support Ctr
Ft Eustis, VA 23602

HART, Fred L
Kinton, Inc
4660 Kenmore Ave
Alexandria, VA 22304

HAWRYSH, Fred, LCdr
National Defense HQ
Canadian Forces Directorate of Military
Occupational Structures
Ottawa, Ontario, Canada KIA OK2

HAZEL, Joe T, Dr
AFHRL/ORE
Brooks AFB, TX 78235

HEAD, Lawrence, W, Jr
USAF School of Applied
Crypotologic Sciences
USAF SACS/TTRE
Goodfellow AFB, TX 76903

HENDERSON, Robert G
Defense Language Inst
Foreign Language Ctr ATFL-TD-TD
Presidio of Monterrey, CA 93940

HENDRIX, William H, Maj
AFHLR/ORS
Brooks AFB, TX 78235

HERMANSEN, Arthur G.
USA Tng Spt Ctr
Ft Eustis, VA 23604

HICKS, Jack M
Army Rsch Inst
5001 Eisenhower Ave
Alexandria, VA 22333

HILLER, Jack H, Dr
US Army Rsch Inst
P.O. Box 5787
Presidio of Monterrey, CA 93940

HITCHO, Karen S
National Security Agency
Ft George G Meade, MD 20755

HOGGATT, Ralph S, Col
AFHRL/NA
Brooks AFB, TX 78235

HOGLE, Janet
Army Education Center
Kisks Hall
Ft Richardson, AK 99505

HOLLAND, J D, Cmd
NODAC
Bldg 150 WNY, Anacostia
Wash D.C. 20370

HOLZ, Robert
US Army Rsch Inst
Atlanta, GA

HOWARD, Charles W, Dr
Kinton, Inc
4660 Kenmore Ave
Alexandria, VA 22304

HUDSON, Ronnie M
Leadership Development Staff
RTC
Yorktown, VA 23690

HUNTER, David R
AFHRL/PE
Brooks AFB, TX 78235

HUPP, Dennis J
Army Personnel Center
Alexandria, VA 22331

ITOGA, Bert K, Maj
AFHRL/XR
Brooks AFB, TX 78235

JACKSON, David K, L/C
AFROTC
4072 Strathmore Dr
Montgomery, AL 36116

JACKSON, Robert A
Northrop Aircraft
Lackland AFB, TX 78235

JACKSON, William L
Directorate for Trng Dev
US Army Aviation Ctr, ATZQ-TD-TAD-TD
Ft Rucker, AL 36362

JACOBS, Louis L
Army Educ Center, Ft Pol, LA
71459

JENNINGS, Dan D
Data Design Lab
1755 S Jefferson Davis Hwy
Arlington, VA 22202

JENSEN, Harald
AFHRL/PE
Brooks AFB, TX 78235

JOHNSON, Clarence A
AFHRL/ORS
Brooks AFB, TX 78235

JOHNSON, Robert N
US Army Administration Center
Ft Benjamin Harrison
Indianapolis, IN 46216

JOHNSON, J. Roger, Capt
OMY
Lackland AFB, TX 78236

JOHNSON, Russell W, L/C
OMY
Lackland AFB, TX 78236

JOHNSON, Tommy J
AFHRL/SM
Brooks AFB, TX 78235

JONES, Jean, Dr
Army Rsch Inst Fld Unit
HQ TCATA ATTN: PERT-OH
Ft Hood, TX 76544

JONES, Karen N
USCG Inst, PO Substation 13
Oklahoma City, OK 73169

KANTOR, Jeffrey E, Dr
AFHRL/PE
Brooks AFB, TX 78235

KEATES, W. E, Capt
Air Command Headquarters
CFB Winnipeg
Westwin, Manitoba, Canada R2R OTO

KEETH, James B
OMY
Lackland AFB, TX 78236

KESSLER, John J, Dr
US Army Rsch Inst
5001 Eisenhower Ave
Alexandria, VA 22333

KING, F J
405 Education Bldg
Florida State University
Tallahassee, FL 32306

KNERR, Clarence S, Dr
US Army Rsch Inst for Behavioral
& Social Sciences
5001 Eisenhower Ave
Alexandria, VA 22333

KNUTSON, Louise C
AFHRL/SM
Brooks AFB, TX 78235

KOHL, Delbert E
Marine Corps Inst
Box 1775
Washington, D. C. 20013

KOPLYAY, Janos B, Dr
AFHRL/SM
Brooks AFB, TX 78235

KUBALA, Albert L
Human Resources Rsch Org
HQ TCATA, ATTN: ATCAT-HUM
Ft Hood, TX 76544

LANG, Donald, Maj
Director Pers Appl Rsch
National Defense Div
Ottawa, Ont K1A OT5

LANTERMAN, Richard S
HQ USCG
Washington, D.C. 20590

LAVENDER, Robert F, Ensign
USCG Training Center
Cape May, NJ

LAVERNE, John E
DTD-USAMPS/TC
Cmdr-US Army Mil Police Sch
ATTN: TDP-C
Ft McClellan, AL 36205

LECZNAR, William B
AFHRL/XR
Brooks AFB, TX 78235

LEECH, Carl A, L/C
National Defense HQ
101 Colby Dr.,
Ottawa, Ontario, Canada
Canadian Forces Directorate of Military
Occupational Structures

LEISEY, Sandra, SrA
AFHRL/PE
Brooks AFB, TX 78235

LIPSCOMB, Suzanne
AFHRL/ORA
Brooks AFB, TX 78235

LOOPER, Larry T
AFHRL/ORS
Brooks AFB, TX 78235

LUNTER, Jeffrey A, A1C
AFHRL/SM
Brooks AFB, TX 78235

MACKENDER, Charles D
Educ Div, G1/DPCA
1st Inf Div
Ft Riley, Kansas 66442

MAGDARZ, Edward F
IBM Corp Dept 804-X585
P.O. Box 12195
Research Triangle Park, NC 27709

MAGNESS, Penny J, Sgt
AFHRL/PE
Brooks AFB, TX 78235

MALONE, Thomas A, Capt
Air University
Sq Officer School Evaluation
42D Area C
Maxwell AFB, AL 36113

MARTIN, Jim
NPRDC
San Diego, CA 92152

MARTINEZ, Raymond
AFHRL/SM
Brooks AFB, TX 78235

MASSAR, R. S.
Testing Directorate, MEPCOM
Ft Sheridan, Il 60037

MASSEY, Randy H, Capt
AFHRL/PE
Brooks AFB, TX 78235

MATTHEWS, John J
AFHRL/PE
Brooks AFB, TX 78235

MATTHEWS, Bettye B
Army Educ Ctr
Ft Richardson, Alaska 99505

McBRIDE, Curtis D
Directorate of Evaluation
US Army Field Artillery School
Ft Sill, Oklahoma 73503

McBRIDE, James R
8506 Lakinhurst Lane
Springfield, VA

McCABE, Jerome M
Data Design Laboratories
1755 So. Jeff Davis Hwy
Arlington, VA 22202

McFARLAND, Barry P, Capt
ASD/ENECC
Wright-Patterson AFB, OH 45433

McFARLANE, Terry, A1C
AFHRL/PE
Brooks AFB, TX 78235

McLANATHAN, F. L, Col
St Mary's University
San Antonio, TX 78284

McLAUGHLIN, Deborah L
Directorate of Trng. US Army
Inf Sch
ATCH-I-B-TMDD
Ft Benning, GA 31905

McMILLIAN, Clyde J, Lieut
Royal Australian Navy
Instructional Design
3813 Leane Dr
Tallahassee, FL 32303

McREYNOLDS, Jane, Dr
AFHRL/DO
Brooks AFB, TX 78235

McVAY, Kenneth W
Missile Munitions Cen & School
ATSK-TD-PM
Redstone Arsenal, AL 35809

MEREDITH, Carlton F, Cdr
USCG Training Center
Governors Island
New York, NY 10004

MEREDITH, John B, Jr
Data Design Laboratories
15 Koger PO 12773
Norfolk, VA 23502

MIRABELLA, Angelo
ARI
5001 Eisenhower Ave
Alexandria, VA 22333

MITCHELL, Jimmy L, Maj
OMY
Lackland AFB, TX 78236

MITZELL, William J, Dr
Trident Training Facility
158 NW Doncee Dr
Bremerton, WA 98310

MOBLEY, Amelia E
US Coast Guard Hq
Recruiting Div, Testing Staff
400 7th St, SW
Wash DC 20590

MOCHARNUK, John B, Dr
McDonnell Douglas Astronautics Co
Box 516
St Louis, MO 63166

MOONAN, William J
NPRDC
San Diego, CA 92152

MOYER, Bobbie E
FLEATSWTRACENPAC
Harbor Dr
San Diego, CA 92147

MULLINS, Cecil J, Dr
AFHRL/PE
Brooks AFB, TX 78235

MURPHY, John W
USAIA
Ft Ben Harrison
Indianapolis, IN 46216

NEAL, William D
SDR Inc
1300 Life of Georgia Tower
Atlanta, Georgia 30308

NELSON, Oliver
HQ ATC/TTSE
Randolph AFB, TX 78148

NEWTON, Tyree H, Col
AFHRL/PE
Brooks AFB, TX 78235

NORDHAUSER, Fred, L/C
AFHRL/NA
Brooks AFB, TX 78235

O'CONNOR, Thomas J, L/C
AFHRL/NA
Brooks AFB, TX 78235

OREND, Richard, Dr
HUMRRO
300 N Wash St
Alexandria, VA 22314

OSBORN, William
HUMRRO
P.O. Box 293
Ft Knox, KY 40121

OSBORNE, James E
Bldg 923 Ellyson
Pensacola, FL 32509

OWEN, Dave
Royal Canadian Mounted Police
250 Tremblam Rd
Ottawa, Ontario, Canada K1A 0R2

PALME, Robert C
NETPDC
Pensacola, FL 32509

PALMER, R. L, Dr
Army Rsch Inst
HQ TCATA (PERI-OH)
W. Fort Hood, TX 76544

PAPPAS, Linda
President's Commission on
Military Compensation (Staff)
Fairfax, VA 22030

PEPPER, von Kluck D
Human Resources Mgt
Naval Educ & Trng Progr Dev Cen
Ellyson, FL 32509

PIPKIN, C M
NETPDC
Pensacola, FL 32509

PINE, Steven, Dr
University of Minnesota
Dept of Psychology
Minneapolis, MN 55455

POLK, Stan, Capt
AFHLR/ORS
Brooks AFB, TX 78235

PORTER, George V
USAF Academy
1510 Hollyhouk Dr
Colo Springs, CO 80907

PORTHOUSE, Harry W, Capt
US Army Trng Dev Inst
Bldg 1514
Ft Eustis, VA 23604

POTEMPA, Kenneth W
AFHRL/XR
Brooks AFB, TX 78235

POWELL, Patrick W, SSgt
AFHRL/SM
Brooks AFB, TX 78235

PRESCOTT, Stanley E, SrA
AFHRL/SM
Brooks AFB, TX 78235

PULLIAM, Robert, Dr
Kinton, Inc
Suite 508, 4660 Kenmore Ave
Alexandria, VA 22304

QUEBE, John C, Capt
AFHRL/PE
Brooks AFB, TX 78235

RASTALL, George, D
CNET CODE N-101
NAS
Pensacola, FL 32508

RATLIFF, Forrest R, L/C
AFHRL/PE
Brooks AFB, TX 78235

RAYNER, Gail T, Dr
Center for Ed Tech
Florida State Univ
Tallahassee, FL 32308

RAYNOR, Bob
Army Education Center
Ft Bragg, NC 28307

REE, Malcolm J, Dr
AFHRL/PE
Brooks AFB, TX 78235

REED, Charles L, Maj
US Army Air Defense Sch
ATSA-TD-IT
Ft Bliss, TX 79916

RICHEY, Brent A, A1C
AFHRL/SM
Brooks AFB, TX 78235

ROBERTS, Wesley K, Dr
Courseware, Inc.
9820 Willow Creek Rd
San Diego, CA 92131

RODGERS, M. N, Capt
Military Leadership & Management Dept
Royal Military College
Kingston, Ontario, Canada

RUBRIGHT, Earl
University of Virginia
164 Rugby
Charlottesville, VA 22903

RUCK, Hendrick
AFHRL/ORA
Brooks AFB, TX 78235

RUDNER, Lawrence M
Gallaudet College-MSSD, Box 113
Washington, D.C. 20002

RUSH, John R
NAVEDTRAP RODEVCEN
Code PD-6
Pensacola, FL 32509

RZUCIDLO, Deborah L, A1C
AFHRL/SM
Brooks AFB, TX 78235

SAKO, Sydney
Officer Training School, USAF
OTS/MTLM
Lackland AFB, TX 78236

SAKUMA, Leslie I
US Army Support Command
Army Education Center
Schofield Barracks Hawaii 96857

SANDERSON, John A
University of Virginia
Dept of Foundations of Educ
Charlottsville, VA 22903

SANDS, William A
Navy Personnel Rsch & Dev Ctr
San Diego, CA 92152

SCANLAND, F. Worth, Dr
Rsch & Program Dev
CNET
NAS Pensacola, FL 32508

SEIBEL, David
D E Seibel & Assoc, Ltd
Box 204
Almonte, Ontario, Canada KOA 1A0

SEUBERLICH, Hans E Oberst
Leiter Gruppe Innere Fuhrung/
Personal
5300 Bonn, Rosenburg

SHAFER, Joe R, L/C
3507 Airman Class Sqd
Lackland AFB, TX 78236

SHEPHERD, Donald O, Dr
Loyola University of Chicago
Chicago, IL 60611

SHIVELY, Albert E
NAVEDTRAPRODEVCEN
Pensacola, FL 32509

SHOEN, William R
Service School Command, NTC
Orlando, FL 32813

SIEGFRIED, Charles W, WO1
Tng Div USA Facilities Engineer
Support Agency
Ft Belvoir, VA 22060

SIMPSON, Arnold Harold
Extension Course Institute
ECI/EDV
Gunter AFB, AL 36118

SKINNER, Jacobina
AFHRL/ORE
Brooks AFB, TX 78235

SLATER, Stephen J, Dr
NW Regional Educ Lab
710 SW 2d St
Portland, OR 97204

SMITH, Judith J, A1C
AFHRL/SM
Brooks AFB, TX 78235

SOLOMON, Elberta
Naval Ed Trng Prog Dev Ctr
Ellyson Fld
Pensacola, FL 32509

TRATTNER, Marvin H
US Civil Ser Commission
1900 E St, NW
Wash DC 20415

TURLEY, L. A., LCDR
CNET
NAS Pensacola, FL 32509

TURNER, James A, Col
OMY
Lackland AFB, TX 78236

VALENTINE, Lonnie D. Jr, Dr
AFHRL/PE
Brooks AFB, TX 78235

VAN RIJN, Paul, Dr
US Civil Service Commission PRDC
1900 E St, NW
Wash DC 20415

VARNEY, Nick C, Capt
Human Factors Testing &
Simulation Rsch
AFHRL/FT
Williams AFB, AZ 85224

VAUGHAN, David, Capt
OMY
Lackland AFB, TX 78236

VINEBERG, Robert, Dr
Human Resources Rsch Organ
27357 Berwick Dr
Carmel, CA 93923

VITOLA, Bart M
AFHRL/PE
Brooks AFB, TX 78235

VOGEL, James A, Dr
Director, Exercise Physiology Div
US Army Rsch Inst of Environmental
Medicine
Natick, MA 01760

WARD, Joe H, Dr
AFHRL/CCS
Brooks AFB, TX 78235

WATSON, Tom
AFHRL/PE
Brooks AFB, TX 78235

WEEKS, Joseph Lloyd
AFHRL/PE
Brooks AFB, TX 78235

WEISSMULLER, Johnny J
AFHRL/SM
Brooks AFB, TX 78235

WELLS, W. L.
AFHRL/SM
Brooks AFB, TX 78235

WILBOURN, James M
AFHRL/PE
Brooks AFB, TX 78235

WILLIAMS, Robert J, Maj
AFHRL/NA
Brooks AFB, TX 78235

YORK, William J, Jr
US Army Signal Sch
Ft Gordon, GA 30905

YOUTT, Robert
US Army Educ Ctr
Ft Devens, Mass 01720

WALDKOETTER, RO, Dr
US Army Rsch Institute
Ft Sill Fld Unit
Ft Sill OK 63503

WALKER, Robert W, Capt
Canadian Forces Pers Applied Rsch Unit
1107 Avenue Rd
Toronto, Ontario, Canada M5N 2E4

WARM, Thomas A
US Coast Guard Inst
P.O. Substation 18
Oklahoma City, Ok 73196

WATERS, Brian K, Maj
AFHRL/TT
Lowry AFB, CO 80230

WEINER, Joseph Co
Army Educ Center
Ft Sam Houston, Texas 78234

WETROGAN, Leon, Dr
US Civil Service Comm
1900 E St, NW
Wash DC 20415

WEYBREW, Ben B, Dr
US Navy Submarine Med Rsch Lab
Box 900 USN Submarine Base
Groton, Conn 06340

WILFONG, Harry D, Dr.
Military Enlistment Processing
Command
Ft Sheridan, ILL 60037

WILLIAMS, Rayburn A
Chief of Naval Educ & Trng
Pensacola, FL 32508

WILLING, Richard C
USCG, Coast Guard Inst
Substation 18
Okla City, Okla 73269

WISKOFF, Martin, Dr
Navy Personnel Rsch & Dev Ctr
San Diego, CA 92152

WOLFINGTON, Robert J
US Army Aviation Ctr
Directorate of Trng Dev
Ft Rucker, AL 36362

WOOD, Adele H, Dr
Naval Health Sci Educ Trng Cmd
National Navy Med Ctr
Bethesda, MD 60024

WORSTINE, Darrel
US Army Military Personnel Ctr
2461 Eisenhower Ave
Alexandria, VA 22331